

**NASA TECHNICAL
MEMORANDUM**

NASA TM X-62,150

NASA TM X-62,150

MATHEMATICS OF PROFILE INVERSION

**Proceedings of a Workshop
held at
Ames Research Center
Moffett Field, Calif. 94035
July 12-16, 1971**

**Lawrence Colin
Editor and Chairman**

August 1972

(NASA-TM-X-62150) MATHEMATICS OF PROFILE
INVERSION L. COLIN (NASA) Aug. 1972
630 p CSCL 12A

G3/19



PRECEDING PAGE BLANK NOT FILLED

PREFACE

This volume contains the technical proceedings of the Workshop on the Mathematics of Profile Inversion, which was held at the NASA/Ames Research Center, Moffett Field, California, during July 12 through July 16, 1971. Ninety-six invited scientists from nine countries representing several scientific disciplines participated in the workshop. Their names and affiliations are given in an appendix.

To better understand the purpose of the workshop and the contents of the proceedings, it is desirable to define broadly, what is meant here by *profile inversion*. In many areas of physical science, fundamental functions that define or describe a physical medium are computed from experimental data obtained through "remote sensing" of that medium, as opposed to direct, or *in situ*, measurement. The fundamental function is related to the experimental data through a differential equation (partial or ordinary), and the reconstruction of the basic function or some of its properties constitutes the general class of inverse problems. The term *inverse* distinguishes these problems from the "direct" problem where, given the physical medium — its fundamental function — one calculates hypothetical experimental results. The direct problem is normally more amenable to solution than the inverse problem, due mainly to the absence of both data limitations and experimental error in the former. In most inversion applications it is important to deduce the spatial variation (radial, height, etc.) of the fundamental function from the experimental data; hence, the inclusion of the term *profile*.

Some time ago, Dr. Mario Grossi of Raytheon Company was pursuing theoretical research concerned with the inversion of radio data to determine the radial refractivity profile of an atmosphere through which the radio waves had propagated — the radio-occultation experiment, now performed successfully on several U.S. planetary probes. In this experiment, coherent radio waves are transmitted from the spacecraft toward earth and recorded there. As the waves traverse the planetary atmosphere and ionosphere they are dispersed (and scattered and absorbed) by those media. The inversion problem is to compute the refractive index distribution (hence pressure, temperature, and electron density) as a function of radius from the center of the planet using the recorded data along with proper algorithms.

Inversion algorithms for this experiment had been under development for several years by the Stanford University and Jet Propulsion Laboratory scientists involved, but a very significant contribution was published in the March 1968 *Journal of Geophysical Research* by two *seismologists*: R. A. Phinney of Princeton University and D. L. Anderson of the California Institute of Technology. They showed that the radio-occultation inversion problem is identical, in principle, to the problem of determining the variation of seismic velocities in the earth from the observed travel times of seismic body waves — what seismologists call the *geophysical inverse problem*. In applying the *geophysical inverse problem* algorithms to the radio-occultation inversion problem, Dr. Grossi recognized that in several other branches of physics, inversion techniques are mandatory for determination of the intended scientific results. He subsequently suggested to the National Aeronautics and Space Administration that all scientists concerned could mutually benefit from a symposium at which the various individual inversion techniques were outlined, discussed, and explored with respect to their interdisciplinary applicability. NASA concurred and requested the undersigned to organize the undertaking.

To ensure adequate representation from each of the major scientific disciplines involved, a Steering Group was formed: M. D. Grossi, Raytheon; J. E. Jackson, Goddard Space Flight Center; H. E. Moses, Air Force Cambridge Research Labs; R. G. Newton, Indiana University; W. Nordberg, Goddard Space Flight Center; A. M. Peterson, Stanford University; R. A. Phinney, Princeton University; and J. Shmoys, Polytechnic Institute of Brooklyn. Each was responsible for organizing and chairing a single, half-day session devoted to his specialty. This responsibility included specifying speakers and their contributions, as well as recommending other participants. Their efforts led to seven formal sessions devoted to the following disciplines: passive atmospheric sounding, active atmospheric sounding, occultation measurements, ionospheric sounding, particle scattering, electromagnetic scattering, and seismology. Some 37 invited papers were presented and discussed. These formal sessions were followed by two informal workshop sessions devoted to short voluntary contributions and a panel discussion exploring, in depth, the interdisciplinary aspects of the various algorithms.

For lack of time, several scientific disciplines involving inversion problems were not formally structured into the program. These include atmospheric sounding by sonic booms and stellar occultation, structural mechanics, plasma diagnostics, lunar conductivity measurements, underwater soundings, biomedical applications, and others. However, most of these areas were represented by participants in the audience, several of whom made short contributions during the last sessions.

The organization of these proceedings generally follows that of the workshop. Seven chapters are devoted to each of the major disciplines. Written accounts of the invited papers are contained therein, each followed by edited accounts of the subsequent discussions. Each chapter concludes with a comprehensive bibliography, including the references from the individual papers. Chapter 8 contains certain of the extended voluntary contributions offered during the workshop sessions, as well as other short, miscellaneous, but important, contributions offered during the panel discussions. Chapter 9 contains the workshop summary papers.

The workshop was unique in that neither the scientific methods nor the scientific results were emphasized or critiqued; the mathematical techniques and algorithms leading to those results were. Of course, it was recognized that often the nature and form of the physical profiles lead naturally to the algorithms involved, and thus are introduced invariably in the papers, but the scientific meanings of these profiles are not important

here. In organizing the workshop, we emphasized scientific areas in which profile inversion is an indispensable part of experimentation. However, the inversion algorithms run the gamut from being almost nonexistent to being highly sophisticated. I feel safe in saying that *each* participating scientist benefited from the workshop in terms of the primary goal. Learning of inversion algorithms developed in one discipline that has application in his own. An added bonus, perhaps not clearly recognized in the organizational stages, was that detailed exploration of the algorithms led to conceptual thinking for *planning* future experiments.

As with any undertaking of this scope, there are many people who contributed to the success of the workshop. The members of the Steering Committee, mentioned above, did an outstanding job in organizing and conducting their sessions and in attracting leading scientists. Special thanks are due W. Smith who, substituting for W. Nordberg at the last moment, did an excellent job in chairing the session on passive atmospheric sounding. Dr. P. C. Sabatier should be singled out for his illuminating contributions and his heroic summary paper. Several Ames personnel helped make the Workshop and the publication of this volume quite a bit easier and less frustrating: Zelda Ballantine, C. Duller, R. Maines, W. Walling, C. Frost, Carol Tinling, and B. Balandis. Finally, and certainly not least, Dr. E. R. Schmerling, NASA Headquarters, provided the impetus and funding that made the workshop possible.

Lawrence Colin
Moffett Field, California

August 1972

PRECEDING PAGE BLANK NOT FILMED

TABLE OF CONTENTS

PREFACE	iii	
CHAPTER 1 - PASSIVE ATMOSPHERIC SOUNDING	1-1	
A Review of Some Aspects of Inferring the Ozone Profile by Inversion of Ultraviolet Radiance Measurements <i>Carlton L. Mateer</i>	1-2	✓ 86
Statistical Retrieval Techniques for Sounding the Meteorological Structure of the Atmosphere <i>C. D. Rogers</i>	1-26	✓ 87
A Review of Nonstatistical Techniques for the Estimation of Vertical Atmospheric Structure from Remote Infrared Measurements <i>B. J. Conrath and I. Revah</i>	1-36	✓ 88
Statistical versus Nonstatistical Temperature Inversion Methods <i>W. L. Smith and H. E. Fleming</i>	1-50	✓ 89
Recent Developments in the Inversion by the Method of Relaxation <i>Moustafa T. Chahine</i>	1-57	✓ 90
Limb Radiance Inversion: Iterative Convergence for a Nonlinear Kernel <i>John C. Gille</i>	1-67	✓ 91
Bibliography	1-77	
CHAPTER 2 - ACTIVE ATMOSPHERIC SOUNDING	2-1	
Acoustic Radar Sounding of the Lower Atmosphere <i>L. G. McAllister</i>	2-2	✓ 92
Atmospheric Radar Sounding <i>R. K. Crane</i>	2-13	✓ 93
Fine Structure of the Lower Atmosphere as Seen by High Resolution Radar <i>J. H. Richter, E. E. Gossard, and D. R. Jensen</i>	2-23	✓ 94
Lidar Probing of the Atmosphere: Some Quantitative Aspects <i>R. T. H. Collis and E. E. Uthe</i>	2-31	✓ 95
Bibliography	2-40	
CHAPTER 3 - RADIO-OCCULTATION MEASUREMENTS	3-1	
Current Methods of Radio-Occultation Data Inversion <i>Arvydas J. Kliore</i>	3-2	✓ 96
Errors Incurred in Profile Reconstruction and Methods for Increasing Inversion Accuracies for Occultation Type Measurements <i>S. H. Gross and J. A. Pirraglia</i>	3-18	✓ 97
Profile Inversion in Presence of Ray Bending <i>H. Andrew Wallio and Mario D. Grossi</i>	3-33	✓ 98
Error Analysis in the Mariner 6 and 7 Occultation Experiments <i>Richard W. Stewart and Joseph S. Hogan</i>	3-50	✓ 99

Error Analysis of Bent-Ray Radio Occultation Measurements	<i>R. L. St. Germain</i>	3-60	✓	11600
An Occultation Satellite System for Determining Pressure Levels in the Atmosphere	<i>Steven G. Ungar and Bruce B. Lusignan</i>	3-65	✓	61
Bibliography		3-78		
CHAPTER 4 – IONOSPHERIC SOUNDING		4-1		
The $P'(f)$ to $N(h)$ Inversion Problem in Ionospheric Soundings	<i>John E. Jackson</i>	4-2	✓	2
The Calculation of Electron Density Profiles from Topside Ionograms—Method and Applications	<i>G. E. K. Lockwood</i>	4-15	✓	3
Numerical Methods for Reduction of Topside Ionograms	<i>L. McCulley</i>	4-27	✓	4
The Mathematical-Physical Problem of $P'(f)$ to $N(h)$ Inversion for Anisotropic Media	<i>A. K. Paul</i>	4-37	✓	5
Automatic $N(h, f)$ Profiles of the Ionosphere with a Digital Ionosonde	<i>J. W. Wright, A. R. Laird, E. J. Violette, D. McKinnis, and D. Obitts</i>	4-50	✓	6
Ionospheric Profile Inversion Using Oblique-Incidence Ionograms	<i>D. L. Nielson and T. M. Watt</i>	4-61	✓	7
Bibliography		4-73		
CHAPTER 5 – PARTICLE SCATTERING: CLASSICAL AND QUANTUM MECHANICAL		5-1		8
Introduction to the Inverse Quantum Scattering Problem	<i>Roger G. Newton</i>	5-2	✓	
Review of the Inverse Scattering Problem at Fixed Energy in Quantum Mechanics	<i>Pierre C. Sabatier</i>	5-14	✓	9
Explicit Expressions of the Potential and Its Derivatives at the Origin in Terms of the Scattering Data	<i>F. Calogero</i>	5-33	✓	10
The Inverse Scattering Problem at Fixed Angular Momentum for Nonlocal Separable Interactions	<i>K. Chadan</i>	5-81	✓	11
Bibliography		5-94		
CHAPTER 6 – ELECTROMAGNETIC SCATTERING		6-1		12
The Inverse Scattering Problem for Transmission Lines	<i>Irvin Kay</i>	6-2	✓	
The Inverse Problem for Radiation Scattering	<i>Reese T. Prosser</i>	6-18	✓	13
Diagnostics of Nonradially Stratified Media	<i>J. Shmoy and J. Pirraglia</i>	6-28	✓	14
Electromagnetic Inverse Scattering	<i>Norbert N. Bojarski</i>	6-34	✓	15
Some Special Cases of the Electromagnetic Inverse Problem	<i>V. H. Weston</i>	6-48	✓	16
Bibliography		6-59		

PRECEDING PAGE BLANK NOT FILMED

CHAPTER 7 – SEISMOLOGY	7-1
Monte Carlo Inversion of Seismic Data <i>Ralph A. Wiggins</i>	7-2 ✓ 17
Backus-Gilbert Inversion of Travel Time Data <i>Leonard E. Johnson</i>	7-10 ✓ 18
The Backus-Gilbert Method and Its Application to the Electrical Conductivity Problem <i>Robert L. Parker</i>	7-21 ✓ 19
Culling of Redundant Data <i>Freeman Gilbert</i>	7-29 ✓ 20
Application of a Stochastic Inverse to the Geophysical Inverse Problem <i>Thomas H. Jordan and J. Bernard Munster</i>	7-36 ✓ 21
Bibliography	7-48
CHAPTER 8 – MISCELLANEOUS CONTRIBUTIONS	8-1
Some Methods for Determining the Profile Functions of Inhomogeneous Media <i>R. Mitra, D. H. Schaubert, and M. Mostafavi</i>	8-2 ✓ 22
k-Space Formulation of the n-Dimensional Scattering Problem <i>Norbert N. Bojarski</i>	8-13 ✓ 23
Inversion of Radiation Data in Biophysics <i>Victor Twersky</i>	8-26 ✓ 24
Transmission Line Inversion and Synthesis from the Point of View of Transient Response <i>M. M. Sondhi</i>	8-30 ✓ 25
The Inversion of VLF-LF Sounder Data to Obtain D-Region Electron-Density Distributions <i>C. H. Shellman</i>	8-37 ✓ 26
The Jost-Kohn Inversion Procedure <i>Reese T. Prosser</i>	8-44 ✓ 27
The Location of the Poles of the Scattering Matrix <i>Peter D. Lax and Ralph S. Phillips</i>	8-50 ✓ 28
Miscellaneous Comments	8-56
CHAPTER 9 – WORKSHOP SUMMARY	9-1
Comparative Evolution of the Inverse Problems <i>Pierre C. Sabatier</i>	9-2 ✓ 29
Additional Comments Concerning Inversion Techniques <i>J. Shroofs</i>	9-35 ✓ 30
APPENDIX – CONFERENCE ATTENDEES	A-1

1. PASSIVE ATMOSPHERIC SOUNDING

The subject of this chapter is the ground-based, airborne, or spaceborne measurement of emitted terrestrial (planetary) or reflected solar radiation. Radiance measurements are made as a function of wavelength (visible, ultraviolet, infrared), look-angle, and atmospheric optical depth, and are inverted to yield the vertical structure of several atmospheric parameters, including temperature, ozone, water vapor, and trace constituents. Regardless of the measurement platform and wavelength band, the mathematical problem is identical: inversion of the integral equation of radiative transfer. Sophisticated statistical and nonstatistical, linear and nonlinear inversion techniques are discussed and compared. The difficulties with the inversion problem are both mathematically and numerically oriented, and the ultimate choice of a "best" method for a particular application may not be based on accuracy but is restricted normally by available resources, quantity of experimental data, and a priori knowledge.

The chapter concludes with a comprehensive bibliography comprising references noted in the text as well as other relevant papers.

W. Nordberg organized the session devoted to passive atmospheric sounding; however, W. Smith chaired the session and led the discussion periods.

A REVIEW OF SOME ASPECTS OF INFERRING THE OZONE PROFILE BY INVERSION OF ULTRAVIOLET RADIANCE MEASUREMENTS

Carlton L. Mateer

Atmospheric Environment Service, Toronto, Ontario

N73-11536

ABSTRACT

The mathematical inversion of light-scattering observations to obtain the atmospheric ozone profile is discussed in terms of the filtering properties of the physical and mathematical processes for different spatial scales. Within this context, it is shown that the physical process of scattering, whether in diffuse transmission or in diffuse reflection, acts as a low-pass filter; which transfers large-scale profile information efficiently to radiance observations but very strongly attenuates small- or fine-scale profile information. To avoid domination of the mathematical inversion by the random error of radiance measurements, an equivalent spatial-scale filtering in the inversion procedure is essential. Methods of inversion used on traditional ground-based Umkehr measurements and on satellite measurements, and the accuracy and utility of the inversion results, are reviewed. The available evidence suggests that mathematical inversion to obtain the low-level ozone profile below 25 to 30 km is either inferior to or no better than the statistical estimation of the profile using total ozone as predictor. However, inversion profiles for high-level ozone above 25 to 30 km appear to have moderately good accuracy.

INTRODUCTION

Indirect inferences about atmospheric structure were made nearly a century ago from the unusual twilight optical phenomena observed following the famous Krakatoa volcanic eruption. The existence of stratospheric dust layers was inferred [Russell and Archibald, 1883] from these observations. According to Mitra [1952], the existence of high temperatures in the upper stratosphere was postulated early in the present century from observations of anomalous sound propagation. The first truly quantitative indirect profiling probably came with the introduction in 1926 of radio methods [Mitra, 1952] for determining ionospheric electron density profiles. Several years later, measurements of diffusely transmitted solar ultraviolet radiation were used to infer the main features of the atmospheric ozone profile [Götz et al., 1934]. In their classic paper, Götz et al. gave the first hints about the fundamental problems that exist in the inversion of radiance measurements to obtain information on atmospheric profiles. However, although quantitative profile inversion is at least 40 years old and has undergone very rapid development in the past decade, it remains a rapidly developing area of scientific interest.

This discussion is restricted to the inference of the vertical ozone profile by the mathematical inversion of radiance measurements of solar ultraviolet radiation scattered by the terrestrial atmosphere. The nature of the ozone profile in the lower atmosphere is illustrated in figure 1, which statistically summarizes over 500 direct profile measurements obtained by balloon sounding at Boulder, Colorado [Dütsch, 1966; Dütsch et al., 1970]. The curves are for different total amounts of ozone given in atm-cm. The main features of the ozone profile are the pronounced maximum occurring at 22-24 km, the tendency to a secondary maximum in the lower stratosphere near 13-14 km when the total ozone amount is high, and the approximately exponential decrease of ozone with height above about 30 km.

The inference of the ozone profile from light scattering measurements may be attempted from ground (diffuse transmission) or satellite (diffuse reflection) measurements. In either case, the radiation scattered by the terrestrial atmosphere comes mostly from a fairly well-defined layer, the *effective scattering layer*. The vertical

position of this layer depends on the ozone absorption coefficient at the wavelength under consideration, the zenith angle of the sun, the direction of observation, and the vertical distribution of ozone in the atmosphere [Dave and Mateer, 1967]. For satellite measurements in the nadir direction, these scattering layers are shown in figure 2, which applies to a solar zenith angle of 60° and a total ozone amount of 0.335 atm-cm. The curves show the relative amount of energy backscattered at various levels in the atmosphere for several wavelengths, with the ozone absorption coefficient varying from about 300 atm-cm^{-1} at 2555 \AA to 0.05 atm-cm^{-1} at 3398 \AA . The area between each curve and the y axis is directly proportional to the backscattered radiance observed at the satellite, and all orders of molecular scattering have been included in the calculation. Similar curves apply to the case of diffuse transmission, as illustrated in figure 3, where the vertical "scanning" effect is achieved by allowing the solar zenith angle to vary for 2 or 3 hr near sunset or sunrise. Although the curves in figure 3 were obtained with single scattering calculations, multiple scattering curves exhibit the same kind of scanning effect.

SOME PROPERTIES OF THE INTEGRAL EQUATION

The inference of atmospheric profiles from radiance measurements usually involves the inversion of an integral equation of the form

$$\int K(x,y)f(x)dx = g(y) \quad (1)$$

where the $g(y)$ are radiance measurements at various values of y , $K(x,y)$ is the so-called "kernel," and $f(x)$ is the unknown atmospheric profile. In some cases, the equation may be of the form

$$\int K[f(x); x,y]f(x)dx = g(y) \quad (2)$$

or

$$\int K[f(x); x,y] dx = g(y) \quad (3)$$

In all cases, these are reducible to a matrix equation

$$Af = g \quad (4)$$

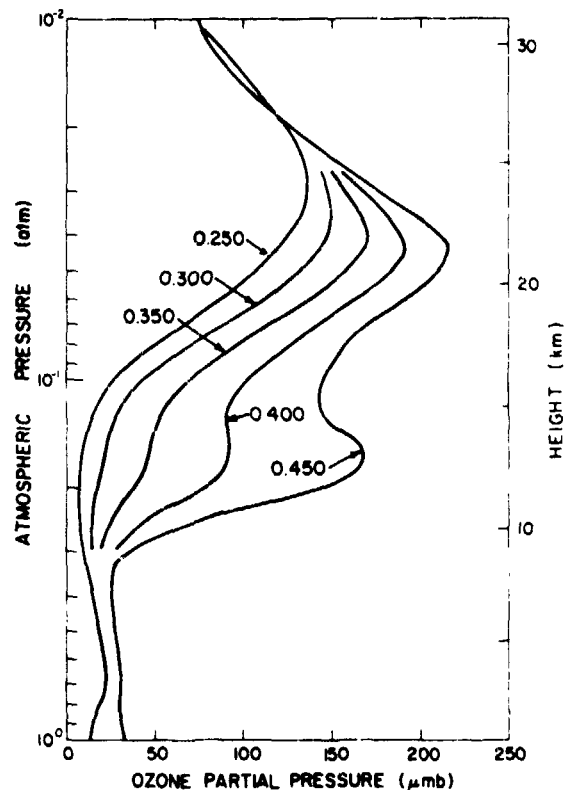


Figure 1. - Average vertical ozone profiles at Boulder, Colorado, for different total ozone amounts.

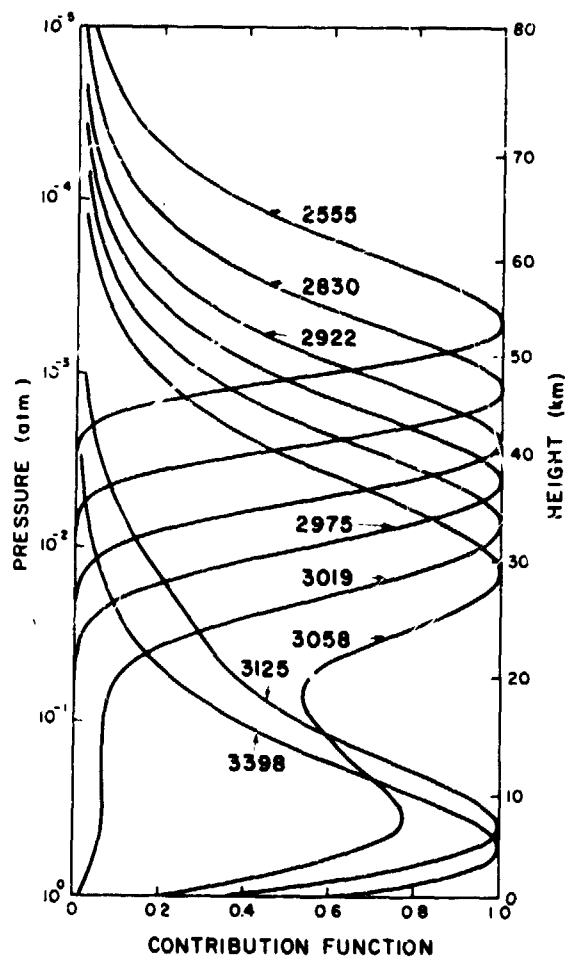


Figure 2.-- Contribution to the nadir-direction radiance by backscattering at various levels in the atmosphere for the case of satellite observations.

or, if $Af_0 = g_0$, $\Delta f = f - f_0$ and $\Delta g = g - g_0$ where f_0, g_0 represent some standard or average conditions, then

$$A\Delta f = \Delta g \quad (5)$$

Throughout this paper, boldface symbols indicate matrixes or vectors; other letter symbols refer to scalars. In profile inversion, the goal is to estimate the unknown profile Δf from radiance measurements Δg , which contain random and nonrandom errors of measurement. Other errors generally inherent in (5) include physical processes that may have been neglected in the mathematical model of the atmosphere described by (1), (2), or (3) and the linearization of what may be a nonlinear equation that is (2) or (3). There may also be small quadrature errors in (5).

The fundamental problems of inverting equations such as (5), when they represent real atmospheric physical processes and involve random errors, have been discussed extensively during the past decade by Twomey [1963, 1965, 1966], Twomey and Howell [1963, 1967], Mateer [1964, 1965], and others. The discussion here is drawn primarily from their work.

For real situations, the problems arise first from the smoothness and large-scale nature of the kernels, directly comparable to the curves of figure 2 for (3); and second from the random errors in (5). In general, the vector Δg will comprise n measurements (say 4 to 12). However, so that the mathematical quadrature will not add appreciably to the random error in

(5), the vector Δf will generally comprise specifications of Δf at $m (> n)$ points (say 20 to 40, but perhaps 100 or more). Immediately, then, the system of equations is underdetermined, admitting an infinite manifold of solutions for Δf that will satisfy (5) exactly. Frequently, this difficulty is avoided by constraining the solution to some ad hoc form so that $m \leq n$ unknowns are sought in the solution. At this point, however, there is no need to artificially restrict m ; indeed, it is convenient to suppose that m is sufficiently large that Δf may adequately describe all features (including the fine structure) commonly found in atmospheric profiles.

It is instructive to think of A as a transformation matrix that transforms from the fx "profile" plane into the gy "observation" plane and to examine the amplitude transmission of different spatial scales from the fx to the gy plane. In doing this, it is convenient to use the eigenvectors that specify the "principal axes" pertinent to the rows of A . We consider the standard matrix algebra equation

$$W^* A^* A W = \Lambda \quad (6)$$

where the (orthonormal) eigenvectors are the columns of W , the asterisk denotes matrix transposition, and Λ is the completely diagonal matrix having the eigenvalues of $A^* A$ on its diagonal and zeroes elsewhere. Note that if $m > n$, there are at most n nonzero eigenvalues and, hence, at most n nontrivial eigenvectors.

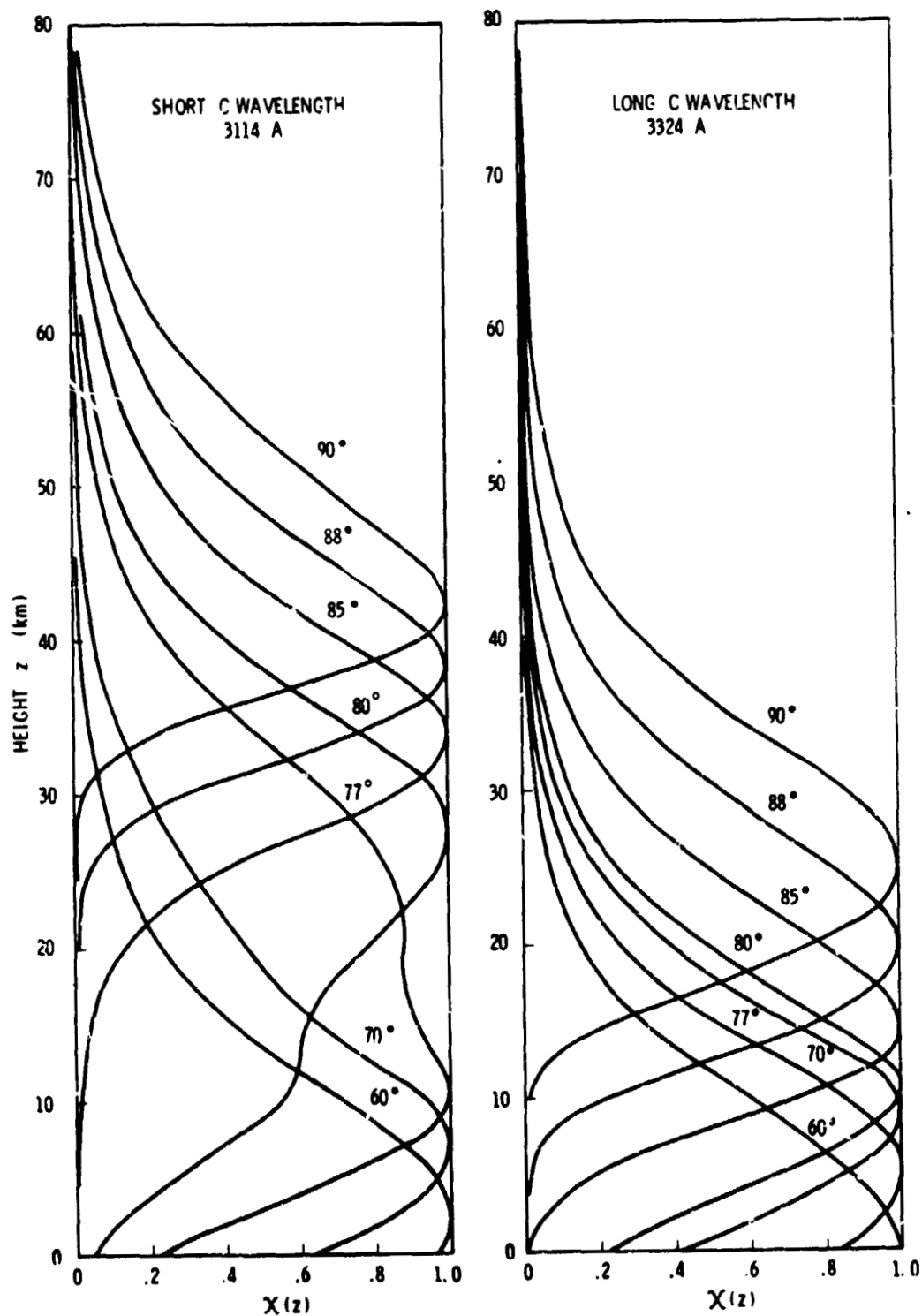


Figure 3.— Contribution $X(z)$ to the zenith-direction radiance by primary scattering at various levels in the atmosphere for the case of ground-based observations [Mareer, 1964].

TABLE 1.— EIGENVALUES AND EIGENVECTORS FOR A TYPICAL MATRIX [after Mateer, 1965]

Vector	Eigenvalue ¹	Eigenvector points								
		1	2	3	4	5	6	7	8	9
1	4.68 (1)	-0.180	-0.471	-0.462	-0.405	-0.321	-0.125	0.136	0.308	0.368
2	1.55 (1)	-0.087	-0.022	0.160	0.262	0.374	0.441	0.515	0.399	0.371
3	4.15 (0)	0.334	0.472	0.248	-0.004	-0.231	-0.356	-0.163	0.334	0.532
4	9.17 (-1)	0.682	0.187	-0.219	-0.297	-0.166	0.227	0.468	-0.020	-0.254
5	1.16 (-1)	-0.493	0.282	0.263	0.045	-0.391	-0.266	0.528	0.096	-0.306
6	9.47 (-3)	0.263	-0.299	-0.082	0.288	-0.347	-0.651	0.180	0.310	-0.283
7	1.58 (-3)	-0.236	0.456	-0.273	-0.489	0.521	-0.031	-0.156	0.311	-0.171
8	3.41 (-4)	-0.122	0.378	-0.698	0.510	-0.064	-0.080	0.118	-0.211	0.159
9	1.03 (-5)	0.009	0.006	-0.115	0.306	-0.356	0.331	-0.354	0.622	-0.388

¹ Number in parentheses is power of 10 by which preceding number is to be multiplied.

We expand the profile Δf in terms of the eigenvectors; thus

$$\Delta f = Wb \quad (7)$$

where b is a column vector of coefficients. This expansion is perfectly general in the sense that Δf can be expanded exactly in terms of m orthonormal vectors. Moreover, the variance $V_f = \sum_{i=1}^m \Delta f_i^2$, of Δf is also given by $\sum_{j=1}^m b_j^2$, since $W^*W = I$, where I is the identity matrix. Table 1 lists eigenvalues and eigenvectors for a case where $m = 9$; the eigenvalues are listed in order of decreasing magnitude while the corresponding eigenvectors are in the appropriate row of the table. Note that the low-order eigenvectors correspond to the large spatial scales (low frequencies) in the profile plane, whereas the higher-order eigenvectors, which exhibit frequent sign changes, correspond to small spatial scales (fine structure, high frequencies). Therefore, the low-order eigenvectors will explain the variance in Δf that arises from the larger spatial scales in the fx plane, while the high-order eigenvectors will explain the variance in Δf that arises from the small spatial scales.

If we now substitute (7) in (5) and calculate V_g , the variance in the gy observation plane, we have

$$V_g = \Delta g^* \Delta g = b^* W^* A^* A W b = b^* \Lambda b = \sum_{i=1}^n \lambda_i b_i^2 \quad (8)$$

where Δg^* , b^* are row vectors. The above result states that the i th spatial scale is amplified by λ_i in the transformation from the profile plane to the observation plane. Thus the fine structure in the profile plane, which is described by the high-order eigenvectors having very small associated eigenvalues, is very strongly reduced in amplitude in the observation plane. Indeed, when $m > n$, the very fine structure in the profile plane will be explained by very high-order eigenvectors for which the λ_i are identically zero ($n < i \leq m$) and will not appear at all in the observation plane. This information about the profile is completely inaccessible, even in n error-free radiance measurements [Twomey and Howell, 1967]. However, for practical purposes, the fine structure information content of the observations is so reduced in amplitude that it is indistinguishable from the random error present in Δg . These results are direct consequences of the smoothness and large scale of the kernel functions of the basic integral equations. As suggested by Twomey [1965], the physical process represented by the transformation matrix A is really a low-pass filter, transmitting low-frequency (large-scale) information quite readily, but providing strong attenuation to high-frequency (fine-structure) information.

There are two obvious consequences of the above result. First, if one examines many observations of Δg , as *Mateer* [1964, 1965] did for Umkehr observations, most of the variance can be explained by large-scale components, suggesting that there were, at most, four linearly independent pieces of information in traditional Umkehr observations. An equivalent result is, for example, that of twelve observations in the same Δg vector, eight could be predicted from the remaining four within the observational error [*Twomey*, 1966].

The second and more important consequence is that when one attempts to recover Δf from observations of Δg by mathematical inversion of (5), the reciprocal of λ_i enters into the picture. Substituting (7) in (5) and solving for \hat{b} by least squares (restricting \hat{b} to $m' \leq n$ elements, where m' is the number of nonzero eigenvalues), we have

$$\hat{b} = \Lambda^{-1} W^* A^* \Delta g \quad (9)$$

where \hat{b} is the estimate of b . Therefore, since $b_i \propto \lambda_i^{-1}$, the fine scale in Δg (i.e., the random error) is greatly amplified in the inversion by reason of the very small high-order eigenvalues, and the random observational errors in Δg dominate the inversion procedure. Clearly, some analog to the natural filtering inherent in (5) is an essential ingredient of any meaningful mathematical inversion method.

An obvious method of achieving this filtering is to restrict \hat{b} to m'' elements, where $\lambda_{m''}$ is the smallest eigenvalue for which the corresponding eigenvector is relatively uncontaminated by measurement errors. It is very nearly equivalent to one of the methods proposed by *Twomey* [1963] wherein the inversion solution is constrained to be a minimum departure from f_0 [see also *Philips*, 1962]. This solution is

$$\hat{\Delta f} = (A^* A + \gamma I)^{-1} A^* \Delta g \quad (10)$$

where γ is a Lagrangian multiplier whose magnitude is determined in practice by trial and error, but is implicitly set by the error in Δg and the natural bounds on Δf . It can easily be shown that the following forms are identical to Eq. (10), namely

$$\hat{\Delta f} = W \hat{b} = W(\Lambda + \gamma I)^{-1} W^* A^* \Delta g \quad (11)$$

TABLE 2.- SOLUTION CONTRIBUTION BY EACH EIGENVECTOR FOR AROSA UMKEHR OF 21 MARCH 1962. TOTAL OZONE IS 0.403 ATM-CM. [*Mateer*, 1965]

Eigen- vector number <i>i</i>	Vector coefficient <i>b_i</i>	Explained variance $\lambda_i b_i^2$	Residual variance	Solution contribution for each eigenvector layer-mean partial pressures ($\mu m b$)								
				1	2	3	4	5	6	7	8	9
Standard distribution			204.4	23	42	84	133	134	95	53	20	7
1	0.225	2.4	202.0	0	-3	-2	-2	-1	0	0	0	0
2	2.145	71.3	130.7	-2	-1	8	10	11	8	7	3	1
3	5.433	122.5	8.2	22	77	32	0	-18	-17	-5	5	4
4	-2.590	6.1	2.1	-21	-14	14	14	6	-5	-7	0	1
5	1.215	0.2	1.9	-7	10	8	-1	-7	-3	4	0	-1
6	0.987	0.002	1.9	3	-9	-2	5	5	-6	1	1	0
7	-15.49	0.4	1.5	44	-212	101	136	-113	4	14	-14	4
8	33.87	0.4	1.1	-50	384	-568	311	-30	-24	24	-21	7
9	-43.79	0.05	1.0	-5	-8	142	-241	218	-130	91	-82	25
Final solution (four vectors)				22	101	136	155	122	81	48	28	12

TABLE 3.- SOLUTION CONTRIBUTION BY EACH EIGENVECTOR USING TWOMEY'S METHOD
WITH $\gamma = 0.5$ FOR AROSA UMKEHR OF 21 MARCH 1962. TOTAL OZONE IS 0.403 ATM-CM
[Mateer, 1965]

Eigen- vector number, i	Vector coefficient, b_i	Explained variance $(\lambda_i + \gamma)b_i^2$	Residual variance	Solution contribution for each eigenvector layer-mean partial pressures (μmb)								
				1	2	3	4	5	6	7	8	9
Standard distribution			204.4	23	42	84	133	134	95	53	20	7
1	0.224	2.4	202.0	0	-3	-2	-2	-1	0	0	0	0
2	2.078	69.1	132.9	-2	1	8	10	11	8	6	2	1
3	4.849	109.3	23.6	19	69	29	0	-16	-15	-5	5	4
4	-1.676	4.0	19.6	-14	-9	9	9	4	-3	-5	0	1
5	.229	.03	19.6	-1	2	1	0	1	0	1	0	0
6	.018	0	19.6	0	0	0	0	0	0	0	0	0
7	-.049	0	19.6	0	-1	0	0	0	0	0	0	0
8	.023	0	19.6	0	0	0	0	0	0	0	0	0
9	.008	0	19.6	0	0	0	0	0	0	0	0	0
Final solution				25	101	129	150	131	85	50	27	13

$$\Delta f = A^*(AA^* + \gamma I)^{-1} \Delta g \quad (12)$$

Equation (11), where $b_i \alpha (\lambda_i + \gamma)^{-1}$, is the filtered analog of (9). It is clear that the magnitude of γ controls the filtering effectiveness of the inversion. We choose γ sufficiently small that the large-scale features of Δf are transmitted essentially undamped, but sufficiently large to severely attenuate the fine-structure features that would otherwise appear in the solution. If we substitute (7) in (11) and make use of (6) we have

$$\hat{b}_i = \frac{b_i \lambda_i}{\lambda_i + \gamma} \quad (13)$$

In the double filtering (the physical filtering) described by (5), followed by the inversion mathematical filtering described by (11), the original vector coefficient b_i is reduced by the factor $(1 + \gamma/\lambda_i)^{-1}$. These methods are illustrated in table 2, which lists results according to Eq. (9), with the final solution taken as the sum of the first four vector contributions, and in table 3, which lists results for Eq. (11) with $\gamma = 0.5$. These tables also show the progressive reduction of the variance V_g of the observations as each vector contribution is added to the solution.

In summary, the inference of atmospheric profiles from radiance observations is not so much a problem of solving n linear equations for m unknowns as it is a problem of establishing the degree of filtering that assures proper transformation of the information content of the observations into profile information, and filtering out and discarding of the spurious contributions from the random noise of measurement. The relative magnitude of m with respect to n is of no consequence in the presence of adequate filtering, except from the point of view of accuracy in (5) and computational economy. Thus (10) should be used if $n > m$ or Eq. (12) if $n < m$; either of these is preferable to Eq. (11) or truncated Eq. (9) because matrix inversion is generally computationally faster than calculation of eigenvalues and vectors. Filtering may be accomplished by the relatively direct methods described above or by indirect methods such as those described by *Chahine* [1968], *Conrath* [1969], and *Smith* [1970].

DIFFUSE TRANSMISSION: GROUND-BASED MEASUREMENTS

The classical example of diffuse transmission is the so-called "Umkehr effect," which was first observed by Götz [1931] and was used by Götz *et al.* [1934] to infer the main features of the ozone profile. It was known at that time that most of the atmospheric ozone must be present in a layer somewhere above the earth's surface because the chemically measured ozone concentrations near the surface were far too low, if the ozone were mixed uniformly through the entire atmosphere, to account for the total ozone amounts determined optically from the measured attenuation of the direct solar beam. However, the precise height or center of gravity of the ozone layer was uncertain.

The Umkehr effect is observed when measurements are made of the zenith sky-light radiances at two wavelengths in the solar ultraviolet with the sun near the horizon. If the ratio of the radiance of the shorter (more strongly absorbed) wavelength to the radiance of the longer (weakly absorbed) wavelength is plotted against the sun's zenith angle, this radiance ratio decreases as zenith angle increases until a minimum is reached for a zenith angle of about 85° (when the wavelengths are 3114 and 3324 Å). As the zenith angle increases further, the ratio increases again. The effect is illustrated in figure 4, where the ordinate is -100 times the common logarithm of the radiance ratio. Götz reasoned that the angular position of the reversal was related to the height of the ozone layer and that it occurred when the effective scattering layer for the shorter wavelength had risen above the ozone layer (see fig. 3).

In obtaining the first indirect profiles, Götz *et al.* [1934] recognized many of the difficulties that have been "rediscovered" in more recent years—namely, the strong interdependence between the points on the Umkehr curve and the fact that the shape of the curve depends mainly on the total amount of ozone. By dividing the atmosphere into rather thick (15 km) layers and assuming constant ozone distribution in each layer, Götz *et al.* were able to obtain a block (histogram) solution through which they later drew a smooth curve so that the amount of ozone in each layer was the same as in the block distribution, but ozone density varied smoothly with height. They correctly inferred that the ozone center of mass was just above 20 km for average midlatitude total ozone amounts, and that the main ozone changes accompanying changes in total ozone occurred in the lower stratosphere so that the center of gravity became lower. Their main result was verified by direct measurement by Regener and Regener [1934], who sent a spectrograph aloft by balloon to 32 km.

Because only the radiance ratio, and not the individual radiances, is available for Umkehr observations, the equation usually used in recent years is [Mateer and Dütsch, 1964]

$$\sum_{i=1}^m \frac{\partial U(\theta)}{\partial x_i} \Delta x_i = N(\theta) - U(\theta) - S(\theta) \quad (14)$$

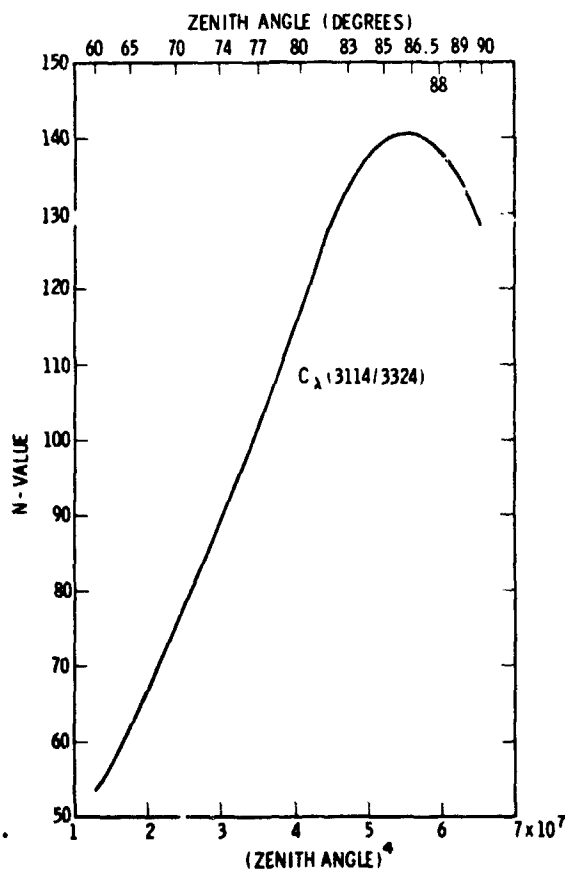


Figure 4.— Umkehr curve for observations at Edmonton, Canada, on the afternoon of May 22, 1961 [Mateer, 1964].

where θ is the sun's zenith angle and

$$N(\theta) = 100 \left[\log (I(\theta)/F_O)_{3324} - \log (I(\theta)/F_O)_{3114} \right] \quad (15)$$

The quantity N may be thought of as an observed relative logarithmic attenuation (having units of centibels). U is the relative logarithmic attenuation for some "standard" ozone distribution, x_i is the amount of ozone in the i th layer of the numerical model of the atmosphere, and $S(\theta)$ represents nonlinear effects ignored in the first-order Taylor expansion. Customarily, measurements are taken while the sun's zenith angle varies between 60° and 90° . The equivalence of Eq. (14) to Eq. (5) is clear: A is the matrix of first-order partial derivatives, the Δx_i are the elements of Δf , and $N(\theta) - U(\theta) - S(\theta)$ comprise the elements of Δg . The equation is solved iteratively,

$$\sum_{i=1}^m \frac{\partial U_j}{\partial x_i} (\Delta x_i)^\nu = N_j - U_j - S_j^{\nu-1} \quad (16)$$

where the superscript ν refers to the iteration, the subscript j refers to the zenith angle, $S_j^0 = 0$, and

$$S_j^\nu = \frac{1}{2} \sum_i \sum_k \frac{\partial^2 U_j}{\partial x_i \partial x_k} (\Delta x_i)^\nu (\Delta x_k)^\nu \quad (17)$$

The iterations are continued until

$$\sum_j |S_j^\nu - S_j^{\nu-1}| \leq \delta \quad (18)$$

where δ is some suitably small number. Further details of this method may be found in *Mateer and Dütsch* [1964].

When solutions of this system are obtained by the filtering methods of Twomey described earlier, they are found to exhibit large-scale differences from concurrent (nearly simultaneous) direct balloon sounding results [*Mateer and Dütsch*, 1964; *Bojkov*, 1966; *Craig et al.*, 1967]. In particular, these Umkehr results show, on the average, too little ozone near the main maximum and too much ozone at lower levels in the atmosphere.

Since the analysis of the scale transmission properties of the inversion procedure indicates that the large-scale features of the profile should come through with good fidelity, one suspects that the errors are introduced by inadequacies in either the physical model of the atmosphere or the treatment of nonlinearities, or both. *DeLuisi* [1969] has shown that a better treatment of the first of these produces only partial improvement. Consequently, one concludes that the nonlinearities must be important and, in particular, that a poor choice for the standard distributions from which the solutions are computed may be responsible.

The optimum statistical inversion method [*Rodgers*, 1966; *Strand and Westwater*, 1968] has been applied to the evaluation of Umkehr observations by *DeLuisi and Mateer* [1971]. The mathematical basis for this method is described in detail by C. D. Rodgers in the next section. DeLuisi and Mateer used Eq. (16) in the form

$$\sum_{i=1}^m \frac{\partial U_j}{\partial \ln x_i} \Delta \ln x_i = N_j - U_j \quad (19)$$

TABLE 4.- VARIANCE OF BOULDER OZONE DISTRIBUTIONS AFTER APPLICATION OF THE TOTAL OZONE REGRESSION. VARIANCE UNITS ARE $(\log_e)^2$ [DeLuisi and Mateer, 1971]

Layer	Layer boundaries, mb	Variance
1	1000-250	0.0814
2	250-177	0.3763
3	177-125	0.2529
4	125-88.4	0.1332
5	88.4-62.5	0.0430
6	62.5-44.2	0.0194
7	44.2-31.2	0.0109
8	31.2-22.1	0.0059
9	22.1-15.6	0.0107
10	15.6-11.0	0.0205
11	11.0-7.81	0.0263
12	7.81-5.52	0.0260
13	5.52-3.91	0.0185
14	3.91-2.76	0.0100
15	2.76-1.95	0.0041
16 ¹	1.95-1.38	0.0008

¹ Ozone amount above the top of layer 16 is 1.7 times the ozone content of layer 16.

ozone in the atmosphere, some assumption had to be made about the ozone profile above 30 km. This was done by smoothly merging each observed profile onto Dutch's [1959] standard profile in such a way that all profiles in the statistical sample were identical above about 45 km. As indicated above, the matrix S_f was calculated from the deviations of each individual profile from the corresponding first-guess profile. The diagonal elements of S_f are listed in table 4 along with the pressure boundaries for each layer in the evaluation model.

where $\ln = \log_e$. They ignored nonlinear effects and did not perform iterations on the premise that a good statistical first guess was used as the standard distribution. Here the optimum solution profile is given by

$$\hat{\Delta f} = S_f A^* (A S_f A^* + S_e)^{-1} \Delta g \quad (20)$$

where S_f is the covariance matrix of the ozone profile (the elements are in effect

$$(S_f)_{ij} = \frac{1}{L} \sum_{k=1}^L \Delta \ln x_{ik} - \Delta \ln x_{jk}, \text{ where}$$

subscript k refers to the k th profile in the sample and $\Delta \ln x_{ik}$ is the logarithmic deviation in the i th layer between the true profile and the first guess profile) and S_e is the covariance matrix for the random errors present in Δg .

The first-guess ozone profile was obtained by a log-linear regression, using total ozone as predictor, and was based on the 500 odd direct balloon soundings at Boulder (fig. 1). Since the direct balloon soundings normally terminate at about 30 km, but the radiance measurements are influenced by all the

TABLE 5.- AVERAGE AND VARIANCE OF EMPIRICAL ADJUSTMENT TO OBSERVED UMKEHR CURVES AT VARIOUS LOCATIONS [DeLuisi and Mateer, 1971]

Solar zenith angles, deg	Arosa		Aspendale		Tallahassee	
	Average, cb	Variance, (cb) ²	Average, cb	Variance, (cb) ²	Average, cb	Variance, (cb) ²
65	-0.10	0.27	()	()	0.10	0.18
70	-0	1.00	()	()	0.07	0.58
74	-0.45	1.95	-0.23	0.44	-0.07	1.12
77	-0.78	3.47	-0.34	0.76	-0.27	1.77
80	-0.91	4.94	-0.31	1.68	0.10	2.44
83	-0.22	8.12	1.14	4.00	1.00	4.73
85	0.96	7.83	2.97	6.48	2.63	5.87
86	1.69	8.01	4.41	8.47	3.96	6.02
88	2.27	8.80	5.48	12.04	4.63	5.22
89	1.93	10.17	5.61	13.58	4.18	4.11
90	.81	11.06	5.06	16.15	3.23	3.22

¹ The smallest zenith angle available for all umkehrrs at Aspendale is 70°. The N-value at 70° is subtracted from the N-values at the remaining zenith angles to remove the instrumental constant.

DeLuisi and Mateer next estimated S_e from samples of nearly simultaneous Umkehr measurements and directly measured ozone profiles at Arosa (Switzerland), Aspendale (Australia), and Tallahassee (Florida). They did this by using each profile and (19) to compute the N_j values that "should" be observed for each profile. The differences between the observed N_j and these computed N_j were considered representative of the effects of errors of measurement not only of the radiance ratios but also of the ozone profiles, and to include errors introduced by inadequacies of the physical model of the atmosphere and the nonlinearities ignored in (19). The average differences $N_j - N_j'$ are listed in table 5 along with the diagonal elements of S_e . Solutions were obtained from (20) after first adjusting the actual observed Umkehr data according to the average $N_j - N_j'$ at each location.

Results are presented in tables 6, 7, and 8 for Arosa, Aspendale, and Tallahassee, respectively. In each table the second column lists the average ozone profile in terms of the mean ozone partial pressure in micromillibars (μmb) in each layer. The third and fifth columns list the average first-guess and solution profiles, respectively. The fourth and sixth columns list rms errors for these profiles. At the bottom of each average column is the average total ozone, while at the bottom of each rms column is the overall rms error, for layers 1 through 11, in terms of $\Delta \ln X_j$, in natural log units. The overall rms error is a measure of the overall fractional error in the layers for which the balloon sounding data were available.

The tables indicate that the optimum statistical method achieves only modest improvement over the first-guess profile obtained from the total ozone regression. Since this modest improvement was attained through the use of an average empirical adjustment vector derived from the dependent sample, the utility of Umkehr observations for deriving *additional* information about the ozone profile below 30 km is seriously in doubt. (When the adjustment vector was not used, there was a slight improvement over the regression profile at Arosa where the overall rms error was reduced from 0.298 to 0.280 and at Aspendale where the reduction was from 0.256 to 0.220, but at Tallahassee there was a trivial increase from 0.305 to 0.306.) The extent to which the residual error in the ozone profile is fine

TABLE 6.- VERTICAL OZONE DISTRIBUTION AT AROSA IN TERMS OF LAYER-MEAN OZONE PARTIAL PRESSURES. SAMPLE SIZE = 49 [DeLuisi and Mateer, 1971]

Layer	Measured average, μmb	Total ozone regression		Solution	
		Average, μmb	RMS, μmb	Average, μmb	RMS, μmb
1	25.5	23.8	7.0	25.9	6.6
2	41.8	32.1	20.3	36.7	17.6
3	59.9	44.1	23.9	45.5	23.8
4	74.1	60.1	21.2	61.5	18.9
5	108.4	101.3	17.9	100.4	14.8
6	139.8	146.3	16.6	143.0	13.5
7	151.6	163.7	16.8	160.4	14.3
8	137.7	150.1	15.5	146.4	12.9
9	114.8	124.1	12.9	119.7	8.8
10	89.2	94.8	11.0	91.3	7.3
11	65.8	68.7	7.9	66.8	5.4
12	46.8	47.6	6.0	46.7	4.6
13	32.1	31.8	4.2	31.4	3.7
14	21.1	21.0	2.0	20.8	1.8
15	13.6	13.6	0.8	13.5	0.7
16	8.7	8.6	0.2	8.6	0.2
<hr/>					
Total ozone: (atm-cm)	.329	.329	-	.327	-
Overall rms error, layers 1-11 (\log_e)	-	-	.298	-	.255

TABLE 7.- VERTICAL OZONE DISTRIBUTION AT ASPENDALE IN TERMS OF LAYER-MEAN OZONE PARTIAL PRESSURES. SAMPLE SIZE = 33 [DeLuigi and Mateer, 1971]

Layer	Measured Average (μmb)	Total ozone regression		Solution	
		Average, (μmb)	RMS, (μmb)	Average, (μmb)	RMS, (μmb)
1	22.9	22.2	3.9	23.2	3.8
2	31.4	23.2	15.7	25.2	11.4
3	37.0	32.8	16.0	33.5	13.7
4	48.2	49.7	13.7	50.5	12.2
5	92.5	92.3	14.0	93.5	15.7
6	138.2	137.5	13.3	138.1	11.9
7	157.0	157.4	11.8	157.5	9.6
8	149.5	148.3	12.2	147.1	11.6
9	126.1	124.2	14.3	121.1	12.7
10	92.4	95.4	12.0	91.8	7.9
11	63.3	69.0	10.4	66.1	7.5
12	43.3	47.7	7.5	45.8	6.0
13	29.4	31.9	4.1	30.8	3.3
14	19.8	21.0	2.0	20.5	1.6
15	13.1	13.6	0.8	13.4	0.6
16	8.5	8.6	0.2	8.6	0.2
Total ozone (atm-cm)					
		.316	.316	—	.315
Overall rms error, layers 1-11 (\log_e)					
		—	.256	—	.205

TABLE 8.- VERTICAL OZONE DISTRIBUTION AT TALLAHASSEE IN TERMS OF LAYER-MEAN OZONE PARTIAL PRESSURES. SAMPLE SIZE = 39 [DeLuigi and Mateer, 1971]

Layer	Measured average, (μmb)	Total ozone regression		Solution	
		Average, (μmb)	RMS, (μmb)	Average, (μmb)	RMS, (μmb)
1	19.8	21.4	8.2	21.1	6.7
2	19.7	19.0	10.4	17.4	10.2
3	24.7	27.4	11.8	25.4	9.7
4	40.4	44.4	14.0	42.9	11.7
5	80.5	87.3	15.5	87.7	13.7
6	135.5	132.6	15.6	134.3	12.3
7	164.8	153.7	20.7	155.6	18.4
8	157.0	147.2	14.2	148.6	13.0
9	125.7	124.3	12.0	124.9	9.2
10	92.7	95.7	14.6	95.5	11.9
11	67.3	69.2	9.5	68.6	7.5
12	46.9	47.8	5.4	47.4	4.5
13	31.4	31.9	2.8	31.7	2.4
14	20.8	21.0	1.3	20.9	1.1
15	13.6	13.6	0.5	13.6	0.5
16	8.6	8.6	0.1	8.6	0.1
Total ozone (atm-cm)					
		.308	.308	—	.307
Overall rms error, layers 1-11 (\log_e)					
		—	.305	—	.269

structure and not accessible to reduction by inversion of Umkehr measurements has not been investigated. The results of *DeLuisi and Mateer* [1971] suggest that it would be more advantageous to invest effort in improving the regression estimation method rather than in further work on the low-level Umkehr inversion. Similar conclusions have been reached by *Sellers and Yarger* [1969] and *Yarger* [1970] for the case of satellite observations. Insofar as the high-level ozone profile is concerned, no direct comparisons have been made between Umkehr results and concurrent rocket data. However, although the high-level ozone profiles derived from Umkehr observations are particularly sensitive to measurement bias [Mateer, 1965] and apparently haze effects can lead to erroneous conclusions about seasonal variation at high levels [DeLuisi and Furukawa, 1970], the high-level Umkehr results are generally consistent with good rocket results. The high-level profile remains the one area of potential utility of carefully made Umkehr observations.

DIFFUSE REFLECTION

Satellite Measurements

The inference of the ozone profile from satellite measurements of the solar ultraviolet radiation backscattered by the earth and its atmosphere is conveniently divided into two subproblems: the inference of the high-level profile above 25 - 30 km and the inference of the low-level profile below this level. The primary reason for this distinction is economic and lies in the complexity of the light-scattering calculations; for the shorter wavelengths that "scan" the atmosphere above the ozone maximum and do not penetrate through the ozone layer, a single-scattering physical model is quite adequate to calculate backscattered intensity to well within 1%. In figure 2, such wavelengths are 2975 Å and shorter. For wavelengths that penetrate the ozone layer and are backscattered appreciably within the troposphere, multiple scattering calculations are essential and the effects of aerosol scattering as well as cloud and ground reflections become quite important. In addition, a considerable body of a priori statistical information about the low-level ozone profile is available, whereas relatively few reliable data are available for the high-level profile. Because of these difficulties, the high-level inversion has received considerable attention over the past decade, while few pioneering spirits have ventured into the low-level inversion problem.

The possibility of deducing the ozone profile from backscattering measurements was first suggested by *Singer and Wentworth* [1957]. The first mathematical examination of the problem was by *Twomey* [1961], who showed that for a single-scattering atmospheric model, the spectral energy distribution of the backscattered radiance was a Laplace transform of the ozone profile, when atmospheric pressure was expressed as an explicit function of the mass of ozone above a given pressure level. This method has been used in the USSR to evaluate measurements of backscattering from the Soviet satellites [Krasnopol'skiy, 1966; Izenas, 1968; Izenas et al. 1969a, b].

The backscattered radiance I in the satellite nadir direction, for solar zenith angle θ and wavelength λ in a plane parallel atmospheric model, is given by

$$I(\lambda, \theta) = F_0(\lambda) (3\beta_\lambda / 16\pi) (1 + \cos^2 \theta) \int_0^1 \exp \left[-(1 + \sec \theta)(\alpha_\lambda X_p + \beta_\lambda p) \right] dp \quad (21)$$

where

- $F_0(\lambda)$ = extraterrestrial solar irradiance
- β_λ = atmospheric scattering coefficient (atm^{-1})
- α_λ = ozone absorption coefficient (atm-cm^{-1})
- X_p = amount of ozone above pressure p (atm) in atm-cm

If we define

$$Q(\lambda, \theta) = \frac{I(\lambda, \theta) 16\pi}{F_o(\lambda) \cdot 3\beta_\lambda \cdot (1 + \cos^2 \theta)} \quad (22)$$

assume $\beta_\lambda p \ll \alpha_\lambda X p$, and set $k = \alpha_\lambda (1 + \sec \theta)$, then

$$Q(k) = \int_0^1 e^{-kXp} dp \quad (23)$$

This equation has the form of (3), wherein the profile appears only in the kernel and not separately in the integrand. Twomey [1961] changes (23) to the form (1) by letting X become the variable of integration, thus

$$Q(k) = \int_0^{X_0} e^{-kX} \frac{dp}{dX} dX + p(o) \quad (24)$$

where X_0 is the total atmospheric ozone content and dp/dX is the inverse of the ozone mixing ratio. For the large values of k associated with $\lambda \leq 2975 \text{ \AA}$, we may let $X_0 \rightarrow \infty$ at the upper limit of integration (fig. 2). The second term on the right-hand side of (24) represents scattering by air molecules entirely above the ozone layer. It then follows directly that

$$Q(k) = p(o) + kL_k(p) \quad (25)$$

where $L_k(p)$ is the Laplace transform of p —that is,

$$L_k(p) = \int_0^\infty e^{-kx} p(x) dx. \quad \text{The inversion problem}$$

reduces to one of finding a suitable analytical form for $Q(k)$ and looking up its Laplace transform. It is necessary, of course, to ensure that the analytical form for $Q(k)$ effectively filters out the random noise in $Q(k)$ [Twomey and Howell, 1963].

Existing direct observational data suggest that the high-level ozone profile is approximately exponential with height [Green, 1964; Rawcliffe and Elliott, 1966], which in an exponential atmosphere is directly equivalent to

$$Xp = Cp^{1/\sigma} \quad (26)$$

where C and σ are constants that specify the profile, σ being the ratio of the ozone scale height to the

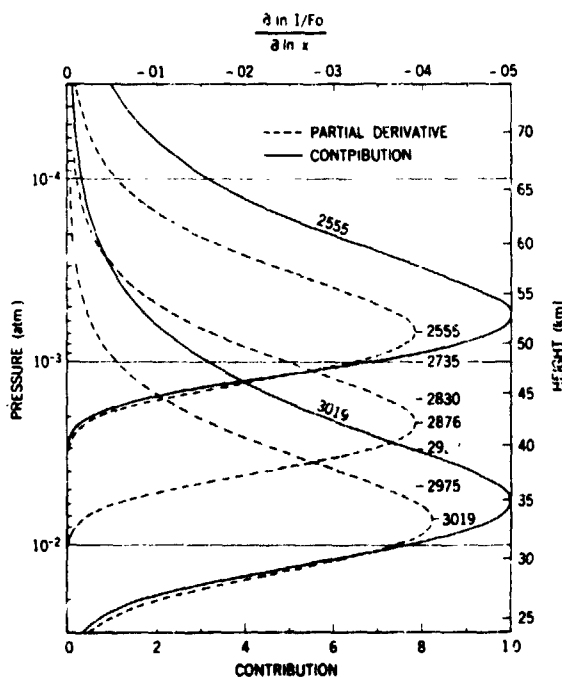


Figure 5.— Contribution to the nadir-direction radiance by back-scattering at various levels in the atmosphere for satellite observations with solar zenith angle 70° (solid curves). Dashed curves are partial derivatives (see also text).

atmospheric scale height. If (26) is substituted in (23) and the integration performed (or solve for p , substitute in (25), and look up the transform), we obtain

$$Q(k) = (kC)^{-\sigma} \Gamma(1 + \sigma) \quad (27)$$

where $\Gamma(x)$ is the gamma function with argument x . According to (27), plotting $\log Q$ against $\log k$ will give a straight line with slope $-\sigma$. Thus a good first approximation to the high-level ozone profile may be obtained by calculating the linear least-squares fit between $\log Q$ and $\log k$.

In attempting to infer *additional* information about the ozone profile from backscattered radiance measurements, it is convenient to substitute (22) in (21), to obtain

$$Q(\lambda, \theta) = \int_0^1 \exp \left[-(1 + \sec \theta) (\alpha_\lambda X_p + \beta_\lambda p) \right] dp \quad (28)$$

Because $Q(\lambda, \theta)$ changes by about one order of magnitude between the shortest and longest wavelength, it is further convenient to scale the system by writing

$$N(\lambda, \theta) = \ln [Q(\lambda, \theta)] \quad (29)$$

and to write the inversion problem as

$$\sum_i \frac{\partial U(\lambda, \theta)}{\partial \ln X_i} \Delta \ln X_i = N(\lambda, \theta) - U(\lambda, \theta) = R(\lambda, \theta) \quad (30)$$

which is similar to (19) in form. Because (30) is a linearization of an essentially nonlinear problem, it is appropriate to perform the inversion iteratively, as follows

$$\sum_i \left[\frac{\partial U(\lambda, \theta)}{\partial \ln X_i} \right]^\nu (\Delta \ln X_i)^{\nu+1} = R^\nu(\lambda, \theta) \quad (31)$$

where ν refers to the ν th iteration. This equation is of the same form as (5); by Twomey's [1963] method, it would have the solution given by the computationally more economical operation of (10) or (12).

The physical situation in (31) is illustrated in figure 5, where the solid curves show the contribution to the radiance in the satellite's nadir direction due to backscattering at various levels in the atmosphere for 2555 and 3019Å. The dashed curves show the partial derivatives for 2555, 2876, and 3019Å, while the short horizontal bars show the position of the partial derivative maximum at intermediate wavelengths. For this case, the solar zenith angle is approximately 75°.

In this section, we examine the properties of solutions obtained by the iterative application of Twomey's method [Smith, 1968; Herman and Yarger, 1969] and by the iterative method described by Smith [1970]. Two

essentially equivalent methods arise from Smith's Procedure. In the first of these, a separate solution is obtained for each wavelength:

$$(\Delta \ln X_k)_{\lambda}^{\nu+1} = \frac{R^{\nu}(\lambda, \theta)}{\sum_i \left[\frac{\partial U(\lambda, \theta)}{\partial \ln X_i} \right]^{\nu}} \quad (32)$$

The solutions for each wavelength are combined to obtain a weighted average solution for each layer, using the partial derivative for each wavelength and layer as the weight, thus

$$(\Delta \ln X_k)^{\nu+1} = \frac{\sum_{\lambda} (\Delta \ln X_k)_{\lambda}^{\nu+1} \left[\frac{\partial U(\lambda, \theta)}{\partial \ln X_k} \right]^{\nu}}{\sum_{\lambda} \left[\frac{\partial U(\lambda, \theta)}{\partial \ln X_k} \right]^{\nu}} \quad (33)$$

The above may be described as the "sum of ratios." In practice, the ratio of the sums gives almost precisely the same result. The latter form is obtained from (30) or (31) by taking $\Delta \ln X_i$ outside the summation and letting $i \rightarrow k$, where k refers to layers for which the partial derivative is large. Each side of the equation is then postmultiplied by $\partial U(\lambda, \theta) / \partial \ln X_k$, summed over λ , and the result is solved for $\Delta \ln X_k$, obtaining

$$(\Delta \ln X_k)^{\nu+1} = \frac{\sum_{\lambda} R^{\nu}(\lambda, \theta) \left[\frac{\partial U(\lambda, \theta)}{\partial \ln X_k} \right]^{\nu}}{\sum_{\lambda} \left\{ \left[\frac{\partial U(\lambda, \theta)}{\partial \ln X_k} \right]^{\nu} \sum_i \left[\frac{\partial U(\lambda, \theta)}{\partial \ln X_i} \right]^{\nu} \right\}} \quad (34)$$

These iterative methods are essentially relaxation methods [Shaw, 1953] wherein it is customary to use an "over-relaxation" factor (F) to speed the convergence, which is usually extremely slow for large-scale forcing functions (the partial derivatives plotted in fig. 5). In the present context, the overrelaxation factor is used in the following manner

$$\begin{aligned} x_i^{\nu+1} &= x_i^{\nu} \exp \left[F(\Delta \ln x_i)^{\nu+1} \right] & (\Delta \ln x_i)^{\nu+1} < 0 \\ x_i^{\nu+1} &= x_i^{\nu} \left[1 + F(\Delta \ln x_i)^{\nu+1} \right] & (\Delta \ln x_i)^{\nu+1} \geq 0 \end{aligned} \quad (35)$$

The first form prevents $x_i^{\nu+1}$ from becoming negative when $(\Delta \ln x_i)^{\nu+1}$ is moderately large and negative, while the second form prevents $x_i^{\nu+1}$ from becoming too large when $(\Delta \ln x_i)^{\nu+1}$ is moderately large and positive. Both forms, of course, give the same result for small $(\Delta \ln x_i)^{\nu+1}$. To determine the best value of F , it is customary to plot the rms residual $\sigma_R = \left\{ 1/n \sum_{\lambda} [R^{\nu+1}(\lambda, \theta)]^2 \right\}^{1/2}$, against F for various numbers of iterations

and to choose the value of F providing the most rapid decrease in σ_R . Such plots are illustrated in figure 6 for values of F between 1 and 2. Although the position of the minimum changes slightly with the number of iterations, the minimum is fairly broad and one would probably select $F = 1.85$ as the best value.

Solutions obtained after 100 iterations with $F = 1.0$, for three different first guesses, are shown in figure 7. The radiance data used here were obtained from the backscattered ultraviolet (BUV) experiment on Nimbus 4. The two straight-line first guesses assume constant ozone mixing ratios throughout the atmosphere and are separated by a factor of 10. The smooth solid first-guess curve is obtained from (27), using least squares to derive σ and C in (26), which is then modified to form [cf. Green, 1964]

$$X_p = \frac{Q_p^{1/\sigma}}{1 + [p/p_*]^{1/\sigma}} \quad (36)$$

which is the same as (26) for small p , but for larger p exhibits a maximum ozone partial pressure at $p = p_*$. For the case illustrated $C = 90.16$, $\sigma = 0.614$, and $p_* = 0.0398$ atm, for p in atm and X_p in atm-cm. In application, p_* is specified from prior knowledge of the low-level ozone distribution. The main purpose of figure 7 is to demonstrate that the solution is essentially independent of the first guess. Solutions obtained from these same first guesses after five iterations by Twomey's method with $\gamma = 0.005$ are illustrated in figure 8. Although the differences between these individual solutions are somewhat greater than those shown in figure 7, they are fairly small between 10^{-3} atm and 10^{-2} atm (1 and 10 μ mb) and the profiles are in good agreement with those in figure 7, between these same pressure levels.

Some other properties of these iterative solutions are illustrated in figures 9, 10, and 11. In figure 9, we show the Twomey solutions, after one, five, and thirty iterations, using $\gamma = 0.005$ and the first guess defined from (36).

The tendency to emphasize the finer structure as the number of iterations increases is quite clear, and γ probably should be increased quite appreciably after the second iteration. What happens in layer 44 (pressure 1.4 μ mb), which is the peak of one of the finer structure features, is shown in figure 10 for Smith's iterative procedure with $F = 1.0$ and 1.85 and for Twomey's method. These curves should be examined simultaneously with the corresponding curves of figure 11, which show the decrease in rms residual (σ_R) with increasing iterations. Two important conclusions may be drawn from these results. First, there is no tendency for the solution to converge in the normal mathematical sense. In this layer, the solution continues to change by a finite amount with each subsequent iteration, while the corresponding decrease in rms residual becomes quite small. The point at which

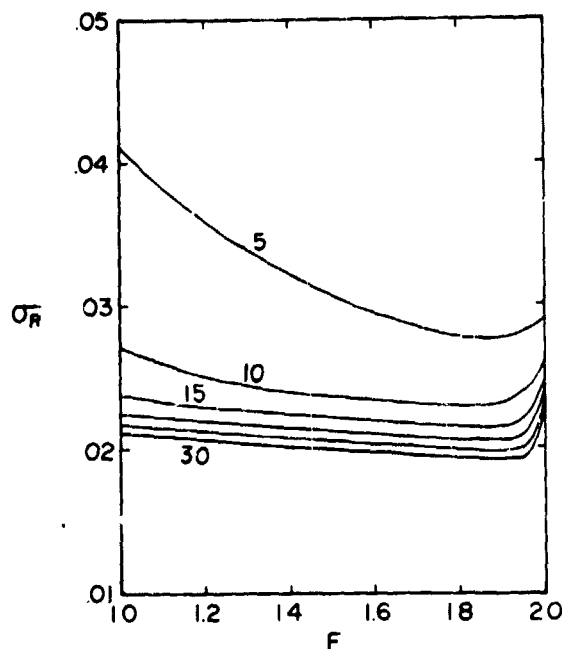


Figure 6.—RMS residual (σ_R) as a function of overrelaxation factor (F) for various numbers of iterations (5, 10, 15, 20, 25, 30) using iterative method of Smith [1970].

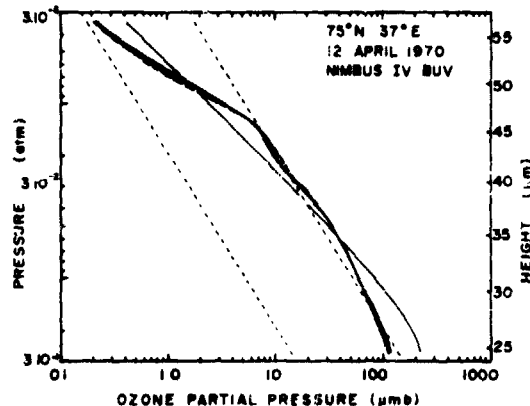


Figure 7.—Inversion ozone profiles (thick lines) obtained after 100 iterations with $F = 1.0$. Corresponding first guesses are shown by thin lines. Smith's iterative method used.

we cease extracting real profile information and begin introducing the noise into the solution profile is by no means clear. The second conclusion to be drawn here is that the most economical way to reach this point is by Twomey's method.

The question of when to terminate the iterative procedure must be geared primarily to the known or assumed standard error of the radiance observations. There appears little reason to proceed beyond iteration ν if

$$(\sigma_R)^\nu \leq \sigma_N \quad (37)$$

or if $|R^\nu(\lambda, \theta)| \leq 2\sigma_N$ for all λ , whichever comes first. Here, σ_N is the standard error of N , the quantity computed from the observations (see (22) and (29)). Since an individual value of N may be in error by more than $2\sigma_N$ in a particular case, it is also useful to have an additional criterion such as

$$|R^{\nu+1}(\lambda, \theta) - R^\nu(\lambda, \theta)| \leq \delta \quad (38)$$

for all λ , where δ is chosen from experience and should be sufficiently small that criteria (37) cause termination of the iterative procedure most of the time.

The accuracy of the iterative inversion procedure is illustrated in figure 12, which shows a rocket profile obtained at Pt. Mugu, California, using an optical rocket-sonde [Krueger, 1969], and two profiles inferred from Nimbus 4 BUUV data [Krueger et al., 1971]. In this case, the inversion has been performed with and without the

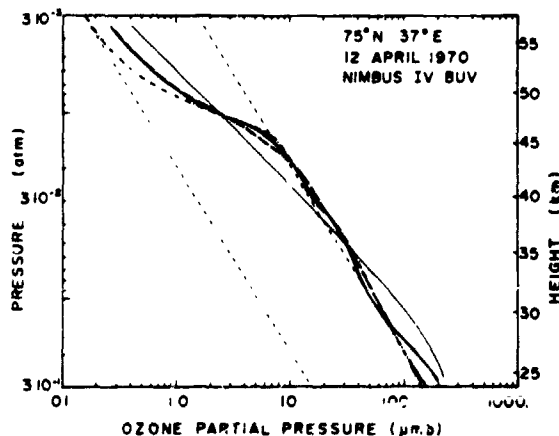


Figure 8. - Inversion ozone profiles (thick lines) obtained after 5 iterations with Twomey's method using $\gamma = 0.005$. Corresponding first guesses are shown by thin lines.

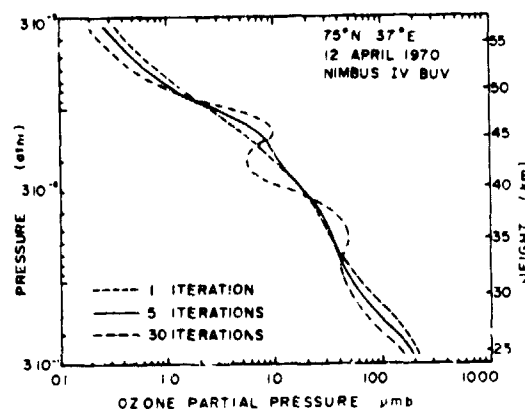


Figure 9. - Inversion ozone profiles after 1, 5, and 30 iterations with Twomey's method using $\gamma = 0.005$. First guess is thin continuous line shown.

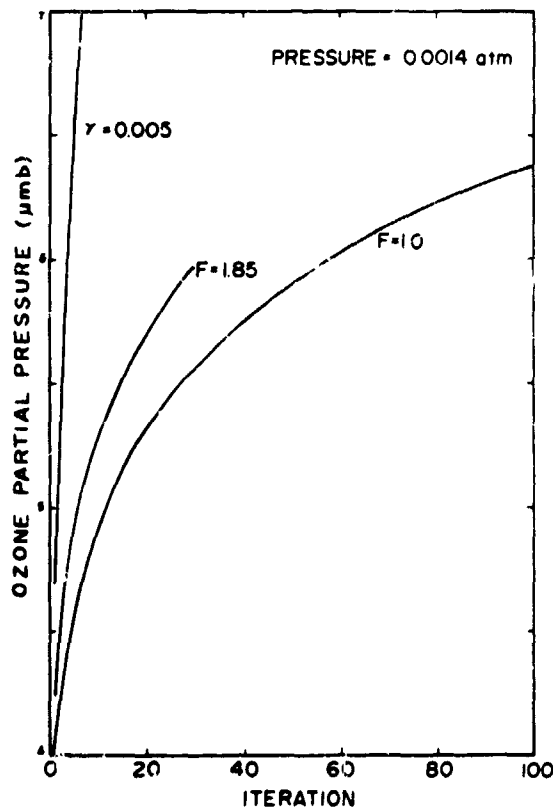


Figure 10. - Mean partial pressure in layer 44 after various numbers of iterations.

2975Å radiance observation. The sun is sufficiently high in the sky for these observations that at 2975Å, there is penetration of the ozone layer and some multiple scattering effect (similar to that exhibited by 3019Å in fig. 2). Consequently, the observed radiance at 2975Å is too high to be consistent with the single scattering model used in the inversion and the profile shows too little ozone near and just above the main maximum. However, the inversion profile obtained without 2975Å is quite good. These satellite profiles have no validity below the ozone maximum.

The Low-Level Ozone Profile

The problem of properly including multiple scattering in the inversion of backscattered radiance measurements to obtain the ozone profile below 30 km was first solved by *Herman and Yarger* [1969]. They considered a satellite experiment in which measurements of polarization components were taken at different observing angles in the plane containing the satellite's nadir direction and the sun. These measurements were taken at a single wavelength. *Yarger* [1970] extended this study by considering the measurements taken by the Nimbus 4 BUUV experiment in which radiance observations are taken at several wavelengths in the satellite's nadir direction. The partial derivative method may also be used here, following the method of *DeLuise and Mateer* [1971] to include multiple scattering in the calculation of the partial derivatives. No attempt has yet been made to obtain the low-level ozone profile by the inversion of real measurements and the present discussion is restricted to results with synthetic measurements. Because the Nimbus 4 BUUV radiance measurements comprise the only available real data suitable to this purpose, the present discussion is restricted to *Yarger's* [1970] results.

In the case of multiple scattering, the equivalent of (21) for the radiance of backscattered radiation in the nadir direction may be written in the form

$$I[\omega(t), \tau_1, \theta, \lambda] = \int_0^{\tau_1} J[\omega(t), t, \theta, \lambda] e^{-t\omega(t)} dt \quad (39)$$

where

t optical depth from the top of the atmosphere ($t = 0$), to some level within the atmosphere, hence defining vertical position in the atmosphere.

$\omega(t)$ albedo of single scattering at level t and is the ratio of scattering optical depth to total (scattering plus absorption) optical depth in unit volume and, hence, defines the ozone profile

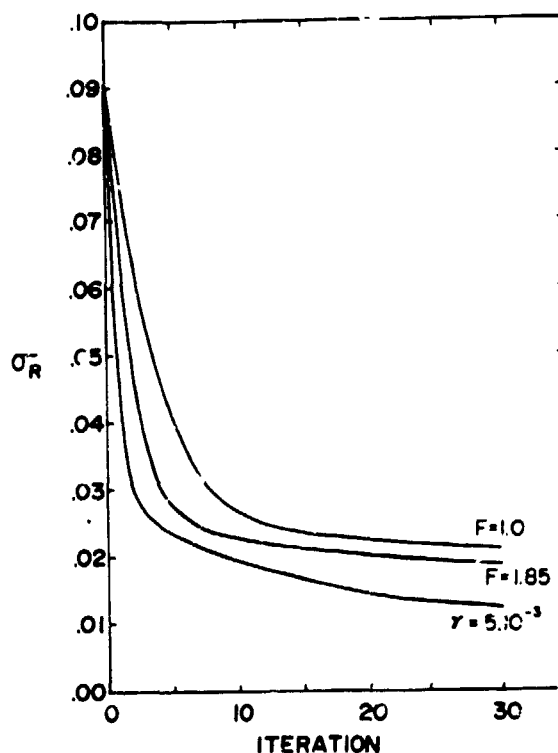


Figure 11. - RMS residual after various numbers of iterations.

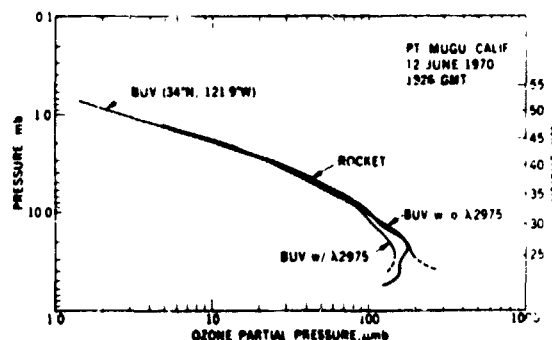


Figure 12. - Comparison of optical rocketsonde and Nimbus 4 (BUV) ozone profiles over Pt. Mugu, California. [*Krueger et al.*, 1971].

τ_1 total optical depth of the entire vertical column of the atmosphere

J the so-called "source function," which is computed in matrix form by successive iteration of the auxiliary equation of radiative transfer [Dave, 1964] and then converted to scalar form for use in Eq. (39); note that J includes the extraterrestrial irradiance, which is shown separately in Eq. (21).

This equation is quite general for the nadir viewing direction in a plane-parallel atmosphere that is horizontally homogeneous and in which only absorption and molecular scattering occur.

Equation (39) has the same form as (2), with the unknown p. profile appearing explicitly in the integrand. However, $\omega(\lambda)$ is wavelength dependent and is not suitable as an unknown for direct solution when we use multiple wavelengths. We have

$$\omega_i(\lambda) = \frac{\beta_\lambda \Delta p_i}{\alpha_\lambda X_i + \beta_\lambda \Delta p_i} = \frac{\beta_\lambda \Delta p_i}{\Delta t_i} \quad (40)$$

where subscript i refers to the i th layer, x_i to the ozone content of the i th layer, Δp_i to the pressure change across the i th layer, and Δt_i the total optical depth of the layer. With this substitution in Eq. (39) and letting $D_i(\lambda) = J_i(\lambda)e^{-\tau_i(\lambda)}$

$$I(\lambda) = \sum_{i=1}^m D_i(\lambda) \beta_\lambda \Delta p_i \quad (41)$$

In this form, (41) is equivalent to (4) with $f = \Delta p$, $A_{i\lambda} = D_i(\lambda) \beta_\lambda$, and $g = I$, in this section I refers to the vector of radiance observations, not the identity matrix. Yarger calls this the pressure increment (PI) method.

In a second procedure, which Yarger calls the ozone increment (OI) method, he proceeds on the v th iteration, with X_i^v , $I^v(\lambda)$, and $D_i^v(\lambda)$, to obtain X_i^{v+1} by writing

$$I(\lambda) = \sum_{i=1}^m [D_i(\lambda) \Delta p_i / \Delta t_i]^v \beta_\lambda \Delta t_i^{v+1} \quad (42)$$

$$I^v(\lambda) = \sum_{i=1}^m [D_i(\lambda) \Delta p_i / \Delta t_i]^v \beta_\lambda \Delta t_i^v \quad (43)$$

Subtracting, he then obtains

$$I(\lambda) - I^v(\lambda) = \sum_{i=1}^m [D_i(\lambda) \Delta p_i / \Delta t_i]^v \beta_\lambda \alpha_\lambda (X_i^{v+1} - X_i^v) \quad (44)$$

which is equivalent to (5), with $\Delta g = I - I^v$, $\Delta f = X^{v+1} - X^v$, and $A_{i\lambda} = [D_i(\lambda) \Delta p_i / \Delta t_i]^v \beta_\lambda \alpha_\lambda$. For his solutions, Yarger uses Twomey's minimum departure from an initial guess in the form shown in (11).

Yarger's inversion results for an ozone profile observed at Boulder, Colorado, on Feb. 20, 1964, are illustrated in figures 13, 14, and 15. Figure 13 shows results for both OI and PI methods when his synthetic measurements contain no random error and his initial guess is derived from a statistical method using total ozone (assumed known) as predictor. Figure 14 shows results for the PI method when his synthetic measurements contain 5% random error and

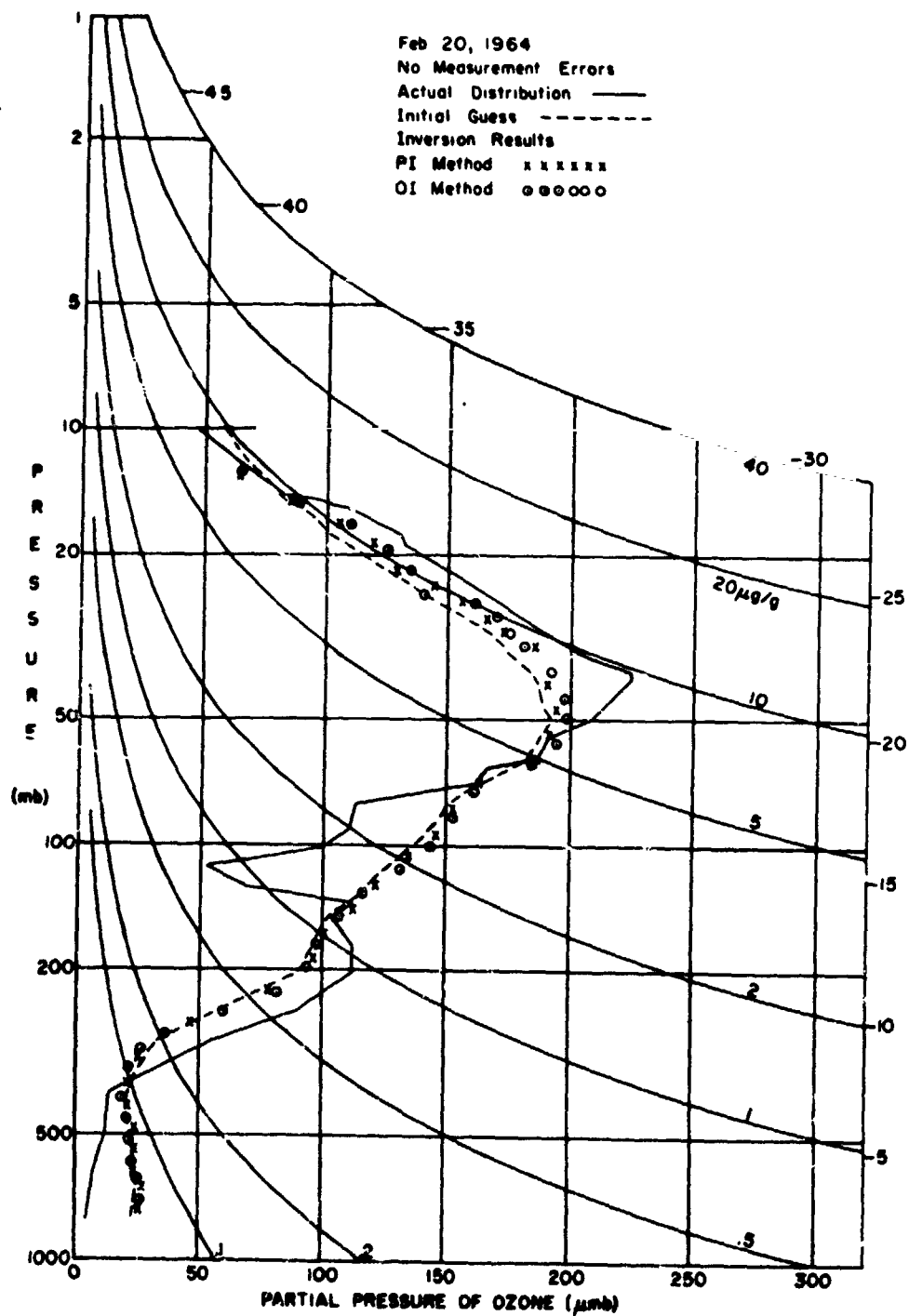


Figure 13.— Inversion, initial guess, and actual ozone profiles for Boulder, Colorado, Feb. 20, 1964, for no measurement errors [Yarger, 1971].

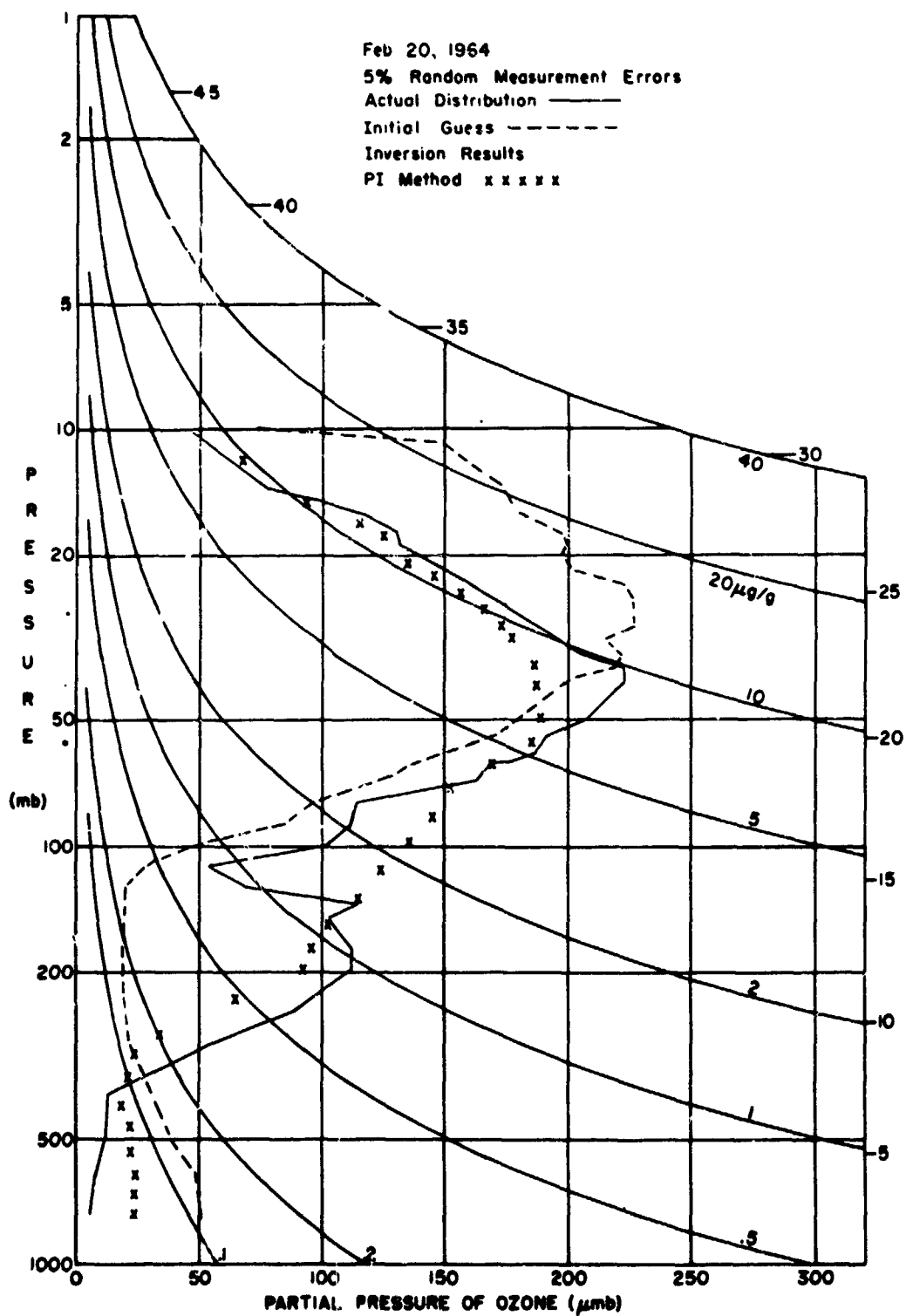


Figure 14. - Inversion, initial guess and actual ozone profiles for Boulder Colorado, Feb. 20, 1964, for poor initial guess with 5% measurement errors. [Yarger, 1971].

he starts from a rather poor initial guess. Apart from a slightly lower partial pressure at the main maximum in figure 14, the solution is essentially the same as that shown in figure 13, suggesting that the result is independent of the first guess. Figure 15 shows Yarger's results for the PI method when the statistical first guess is used and for the two cases: no measurement errors and 5% measurement errors. Because he was unable to achieve any real improvement over his statistical first guess, Yarger concluded, "... there seems no reason to prefer an inversion method based on the inversion of the radiative transfer equation over the statistical method for inferring information concerning the vertical ozone distribution." Since statistical inferences are not without error and because of the result illustrated in figure 14 (poor initial guess), it is difficult to agree completely with his conclusion. However, Yarger's conclusion may be correct for other reasons—namely, the difficulty of properly accounting for surface reflectivity, as well as scattering by haze and clouds in the inversion of real data [Twomey, 1969]. Yarger recognized these additional problems.

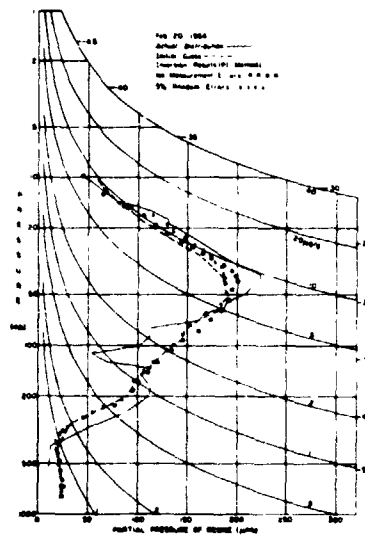


Figure 15.— Inversion, initial guess and actual ozone profiles for Boulder, Colorado, Feb. 20, 1964, for 5% measurement errors [Yarger, 1971].

CONCLUDING REMARKS

It appears that the application of mathematical inversion methods to infer the high-level ozone profile from light scattering measurements provides useful results. However, the inference of the low-level ozone profile remains a venture of somewhat dubious value, not only because of deficiencies in economical computational models, but also because of marginal information retrieval, even under the idealized conditions of synthetic experiments. It may prove more fruitful to measure total ozone by independent techniques and to use statistical methods, with total ozone as predictor, to infer the low-level ozone profile.

From the mathematical point of view, in the absence of prior statistical information, the evidence suggests that the method of Twomey is more economical than the iterative method for inferring high-level ozone profiles from backscattered radiance measurements.

It is quite clear that the mathematical inversion of integral equations of the first kind should not be considered as a solution of a set of linear algebraic equations, but rather as a filtering procedure in which the real information in the radiance observations is transmitted to the solution profile essentially undamped, but the spurious fine structure in the solution profile, which arises from the random noise in the observations (and in the mathematical model), is damped out entirely.

ACKNOWLEDGMENTS

Work on the Nimbus 4 BUV experiment has been carried out in collaboration with my colleagues Dr. Donald F. Heath and Arlin J. Krueger of the Goddard Space Flight Center. I wish to thank Dr. Douglas N. Yarger, Iowa State University, for providing glossy prints of the figures from his paper which appear as figures 13, 14, and 15 in this paper. I am indebted to Dr. Richard Asselin, Atmospheric Environment Service, for discussion of the use of relaxation factors in iterative solutions.

DISCUSSION

Parker: What is the total number of data observations entering in the work where you have so many iterations to solve the equations? Clearly not the six you mentioned earlier.

Mateer: Five, six, or seven, depending on zenith angle of sun.

Parker: Why do you need an iterative technique to solve a bunch of linear equations when there are only six of them?

Mateer: Because the system is not linear. You make a first guess and compute residuals. There are two ways to do this and you should compare them to see if the system is linear.

Bojarski: What is an upper limit of N that you would like to be able to solve? Obviously, the more the better.

Mateer: Not necessarily. As shown by Twomey the measurements are linearly dependent. Suppose that you have 12 observations. Given say the best 4 of those 12, you can predict the other 8 within the error measurement. I would like to have about 6, to allow some redundancy.

Bojarski: How about the number of data points in the final result?

Mateer: In the range 70 to 150 points. The matrix to be inverted is still the size of the number of measurements, 6×6 .

Unidentified speaker: What makes the solutions nonunique? You said there were an infinite number of solutions.

Mateer: The variance amplification factor is responsible in the complete solution. The eigenvalues are so small for the fine scale patterns. You must be careful to filter just the right amount.

Chahine: The solution of an integral equation with fixed limits is very stable and is unique if it exists. The instabilities shown here are due to the fact that the equations were linearized, and linearization is improper. This violates two or three basic mathematical rules. The damping functions in the Twomey method and other methods are means of rectifying the destruction of the integral equation when it was changed from a non-linear system into a linear system of equations. The Method of Relaxation is one means of solving the non-linear system without linearizing it.

STATISTICAL RETRIEVAL TECHNIQUES FOR SOUNDING THE METEOROLOGICAL STRUCTURE OF THE ATMOSPHERE

C. D. Rodgers

Clarendon Laboratory, Oxford University, Oxford, England

N73-11587

ABSTRACT

Statistical retrieval methods for remote sounding are reviewed. Methods are given for constraining an essentially incomplete problem by means of the known statistical behavior of the solution. Information content of the observations and the meteorological structure is discussed. Linear versions of maximum probability and minimum variance methods are given in some detail, and extensions to the nonlinear case are described.

INTRODUCTION

The thermal radiation emitted by the atmosphere depends on the distribution of temperature and absorbing substances, including water vapor and clouds. The form of this dependence is known, so that in principle it is possible to sound the meteorological structure of the atmosphere from remote measurements of radiation, either from the ground or from a satellite.

One of the main problems in interpreting remote soundings of an atmosphere is that there are not enough independent quantities in any one observation to determine the meteorological structure as completely as one would wish. There are an infinite number of valid solutions that agree with the observation within experimental error, thus giving us the problem of choosing the 'best' solution in some sense.

Statistical methods use a priori knowledge of the statistical behavior of the atmosphere as a constraint on the solution. Two basic approaches have been used: In the *maximum probability* method, a probability density function (pdf) is set up to describe the atmosphere, and the chosen solution is the one *most likely* consistent with the observations; the *minimum variance* method finds a relation that obtains the meteorological structure from the observations in such a way that the deviation of the derived structure from the original structure is minimum in a statistical sample.

In what follows, matrix notation is normally used, although suffices are used where necessary to clarify the meaning of an equation. Occasionally the same symbol is used for a vector and for a matrix comprising a statistical sample of such vectors. The distinction should be clear from context.

MODELS FOR THE METEOROLOGICAL STRUCTURE

The meteorological structure of the atmosphere at a particular place can be taken to consist of the vertical temperature, humidity, cloud, and possible ozone profiles. These are all continuous functions of height and cannot be represented mathematically. To be treated at all, they must be approximated in terms of a finite number of parameters. The simplest treatment is a discretization in which values of temperature, humidity, and the like are specified at a number of fixed levels and interpolated between these levels. This is a particular case of a general linear model such as

$$\theta(z) = S_0(z) + \sum_{j=1}^p b_j S_j(z) \quad (1)$$

where b_1, b_2, \dots, b_p are p parameters representing the temperature profile $\theta(z)$ as a linear combination of the functions $S_j(z)$. These functions are arbitrary, except for the restriction that they be independent.

Most representations are inefficient in the sense that even though the functions $S_j(z)$ may be algebraically independent, the coefficients b_j are not statistically independent. This is especially true of a discretization, in which the temperature at any one level is strongly correlated with the temperature at nearby levels. It is clear that the numbers of parameters required to describe a profile to a specified accuracy could be reduced if the parameters were statistically independent. It is also clear that the model functions $S_j(z)$ should also be orthogonal, and for convenience they should be orthonormal. Statistical independence implies that the covariance matrix of the parameters b should be diagonal:

$$\frac{1}{M} \sum_{m=1}^M b_{mj} b_{mk} = \Lambda_{jk} \quad (2)$$

where m indicates membership of a sample of size M , and Λ is a diagonal matrix.

We first discretize Eq. (1) so that we can use the algebra of matrices, rather than the algebra of functions:

$$\theta_{mz} = \bar{\theta}_z + \sum_{j=1}^p b_{mj} S_{jz}$$

we let S_0 be the mean profile $\bar{\theta}$. Orthonormality of S implies that $S^T = S^{-1}$, so we can put

$$(\theta - \bar{\theta}) S^T = b$$

and on substituting this in Eq. (2) we obtain

$$\frac{1}{M} S(\theta - \bar{\theta})^T (\theta - \bar{\theta}) S^T = \Lambda$$

so that the functions S are eigenvectors of the covariance matrix of θ :

$$H = \frac{1}{M} (\theta - \bar{\theta})^T (\theta - \bar{\theta})$$

It can be shown that the eigenvalue Λ_{jj} is the amount of variance "explained" by the term $b_{mj} S_{jz}$ in the expansion of θ , and if the expansion is arranged in decreasing order of Λ_{jj} , it can be truncated at any term, giving the most accurate representation for that number of terms for that statistical sample in a least-squares sense.

These functions are known as *empirical orthogonal functions*, or *characteristic patterns* [Lorentz, 1956; Obhukov, 1960; Grimmer, 1963].

TABLE 1.- EIGENVALUES OF SOME MATRICES

i	H		KHK ^T		K K ^T
	1	2	3	4	
1	0.5886	0.5886	0.8941	0.8941	1.37
2	0.1858	0.7744	0.0852	0.9793	0.44
3	0.1075	0.8819	0.0152	0.9945	0.32
4	0.0422	0.9241	0.0033	0.9978	0.185
5	0.0202	0.9443	0.0018	0.9996	0.090
6	0.0114	0.9557	0.0003	0.9999	0.038
7	0.0101	0.9658	0.0000	1.0000	0.014
8	0.0076	0.9734	0.0000	1.0000	0.0008

In the case of H and KHK^T the eigenvalues have been normalized so that $\sum \lambda_{ii} = 1$. Thus cols. 1 and 3 are the fraction of variance explained by the corresponding eigenvector, and cols. 2 and 4 are the fraction explained if the expansion is truncated.

Table 1 illustrates how well the profile of Planck functions, which is related to temperature, can be approximated with a small number of parameters. The sample is for altitudes of 0 to 70 km for the northern hemisphere. The first characteristic pattern explains nearly 60% of the total variance in the sample; 97% of the variance can be explained with eight patterns and therefore with an eight-parameter model. Thus, fewer observations than expected are required to determine the temperature profile.

The analysis in this section has been in terms of temperature only, but the same method can be applied to all the quantities making up the meteorological structure.

THE RADIATION OBSERVATIONS

The radiance I emitted at the top of an absorbing atmosphere can be written as

$$I(\nu) = \int_g^\infty B(\nu, \theta(z)) \frac{dT}{dz}(z, \infty) dz + B(\nu, \theta(g)) \frac{dT}{dz}(g, \infty) \quad (3)$$

where $B(\nu, \theta)$ is the Planck function at wavelength ν and temperature θ , $T(z, \infty)$ is the transmission of the atmosphere between height z and a satellite, and g is the ground. The transmission will depend on the distribution of absorbing gases, and to a small extent on the temperature profile. This relation is essentially an integral transform of B , the Planck function profile, with kernel dT/dz , the transmission gradient.

If we consider a small range of wavelengths, and a constant distribution of absorber, we may discretise Eq. (3) and obtain the following relation:

$$I_i = \sum_z K_{iz} B_z \quad i = 1, \dots, n \quad (4)$$

where I_i is the radiation at frequency ν_i , B_z is $B(\nu, \theta(z))$ and the matrix K_{iz} is related to the transmission gradient. This simplified form is useful for investigating the nature of the problem, although more detail is required in practice. Figure 1 shows typical set of K_{iz} . The ordinate is z , and the curves are labeled with i . The Planck function is a known function of temperature and can be used to replace temperature in the analysis. The general linear model (1) can be used for B_z in a discretized form:

$$B_z = \bar{B}_z + \sum_{j=1}^p S_{zj} b_j$$

If this is substituted in Eq (4) we get

$$I_i = \bar{I}_i + \sum_{j=1}^p A_{ij} b_j$$

where $\bar{I} = K\bar{B}$ and $A = KS$, giving a linear relation between the observations and the model parameters.

The components of the observation I are not statistically independent partly because of the overlap of the kernels (fig. 1), and partly because of the statistical nature of the atmospheric Planck function profile. To get an idea of how much information there is in a sample of observations, we can transform them into a statistically independent form using the same method as used for the vertical profile of temperature. The covariance matrix of the observations is

$$\frac{1}{M} \sum_{m=1}^M (I_{mi} - \bar{I}_i)(I_{mj} - \bar{I}_j) = \frac{1}{M} K(B - \bar{B})(B - \bar{B})^T K^T = KHK^T$$

where H is the covariance matrix of the Planck function profile B . The set of eigenvalues of KHK^T given in table 1 for a typical case with eight kernels shows that over 99% of the variance is explained by the first three characteristic patterns.

The interdependence of the kernels can be expressed by a similar analysis. If we seek a transform L so that the transformed kernels LK are orthogonal, we must have

$$LKK^T L = A$$

when A is diagonal. The required transform is, therefore, the eigenvectors of KK^T , and A_{ii} can be interpreted as the square of the ratio of the transformed signal to the untransformed signal where noise is constant. The number of independent observations can be viewed as the number of eigenvalues that are larger than (noise variance)/(signal variance).

THE LINEAR CASE

The linear case is convenient for illustrating statistical retrieval techniques without the algebraic complications of the general case. The problem is as follows.

Given a set of observations I , which are related to the profile parameters b by

$$I = Ab + \epsilon \quad (5)$$

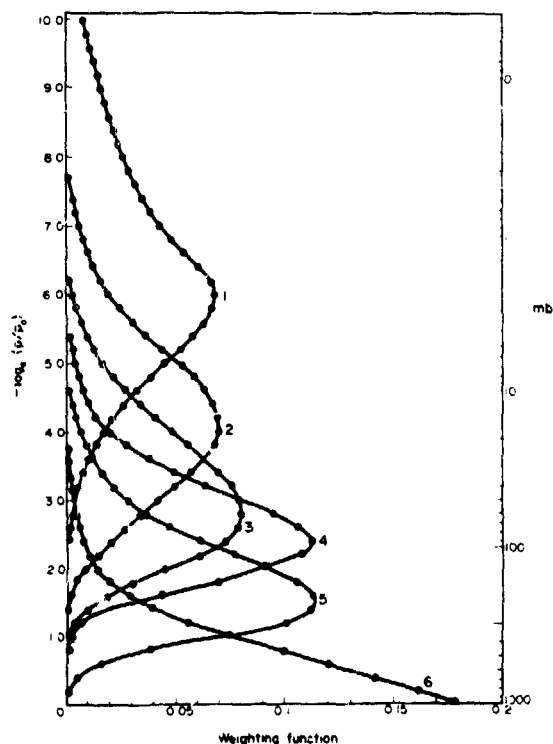


Figure 1.— A typical kernel for sounding the temperature profile, using the CO₂ 15μ band [Rodgers, 1971]

where A is a given matrix and ϵ is a random variable (experimental error) of known statistical behavior, and given a sufficiently large statistical sample of profiles b , find a statistically optimum inverse relation giving b explicitly as a function of I .

Maximum Probability

The maximum probability method finds the profile b that maximizes the conditional probability density function (pdf) $P(b|I)$ —that is, the most likely solution consistent with the observations. The expression $P(b|I)db$ denotes the probability that b lies in the range $b, b + db$ when I is given. Using Bayes theorem, we write:

$$P(I|b) = \frac{P(I|b)P(b)}{P(I)} \quad (6)$$

which is the condition pdf of the observation I given the atmospheric state b . If there were no experimental noise it would be a delta function at $I = Ab$, the theoretical value of the observation. The presence of noise turns it into a gaussian distribution centered at the same place:

$$-2 \log_e P(b|I) = (I - Ab)^T E^{-1} (I - Ab) + \text{const}$$

where E is the covariance matrix of experimental error, which is usually diagonal.

We estimate $P(b)$ from the given statistical sample of b , and if we choose sufficiently restricted regions of time and space, a gaussian distribution can be used:

$$-2 \log_e P(b) = (b - \bar{b})^T H^{-1} (b - \bar{b}) + \text{const}$$

where \bar{b} is a mean and H is the covariance of b about this mean. If we are using an empirical orthogonal function representation of the profile, \bar{b} will be zero and H will be diagonal.

The term $P(I)$ can be ignored in Eq. (5), as we wish to maximize $P(I)$ with respect to b only. This is done by minimizing $-2 \log_e P(b|I)$ with respect to b :

$$\frac{\partial}{\partial b} \left[(b - \bar{b})^T H^{-1} (b - \bar{b}) + (I - Ab)^T E^{-1} (I - Ab) \right] = 0$$

leading to the solution

$$b = (H^{-1} + A^T E^{-1} A)^{-1} A^T E^{-1} (I - \bar{I}) + \bar{b} \quad (7)$$

The most likely profile is thus a linear function of the observations, and the form of this linear function can be precomputed. An alternative form can be found by manipulation:

$$b = HA^T(AHA^T + E)^{-1}(I - \bar{I}) + \bar{b} \quad (8)$$

Note that we do not require that the number of parameters p in the model for the meteorological structure b be less than or equal to the number of measurements n comprising an observation I . The use of statistics has introduced enough extra equations into the system to prevent the problem being incomplete. Of the two forms (7) and (8) of the solution, (7) will be used if $p < n$, and (8) will be used if $p > n$, so that we operate on the smallest matrices. If Eq. (8) is used, there is little need to reduce the number of model parameters; a simple discretization can be used.

Many nonstatistical inverse methods lead to relations of the general form of (7) or (3), allowing us to give a statistical interpretation of the nonstatistical constraints used [Rodgers, 1971].

Minimum Variance

In the minimum variance approach we choose a linear solution b^* of the form

$$b^* = D(I - \bar{I}) + \bar{b}$$

in such a way that the residual variance

$$\frac{1}{M} \sum_{m=1}^m (b - b^*)^2$$

is minimized for the statistical sample. This is straightforward multiple regression of b on I . The solution is

$$D = (b - \bar{b})(I - \bar{I})^T [(I - \bar{I})(I - \bar{I})^T]^{-1} \quad (9)$$

The statistical sample can be obtained by collecting coincidences of remote soundings and direct soundings [Smith, 1969]. This procedure has the advantage that the kernels need not be known accurately, nor is absolute calibration of the instrument necessary. Alternatively, the sample can consist of direct soundings, for which the radiances I are computed using Eq. (5). If we substitute (5) in (8) we get

$$D = HA^T(AHA^T + E)^{-1}$$

where H and E are the covariance matrices defined in the preceding section. The final solution b^* is, therefore, identical to the maximum probability solution

$$b^* = HA^T(AHA^T + E)^{-1}(I - \bar{I}) + \bar{b} \quad (10)$$

This equivalence occurs in the case of gaussian statistics and a linear solution.

Note that although D is in a sense an inverse of A , it is not a generalized inverse in the sense of Moore [1935] and Penrose [1955], because neither DA or AD is a unit matrix. If $E = 0$, then D is a generalized inverse of A . If $E \neq 0$, the solution found from (10) does not give the same radiances as measured, but gives radiances within experimental error of those measured.

The maximum probability method can be extended to include any kind of constraint that can be expressed as a pdf. As described, the statistics could be climatological, covering some period of time and space. However, we could make use of a forecast profile, in which case we use the statistics of forecast errors:

$$-2 \log_e P(b) = (b - b_f)^T H_f^{-1} (b - b_f)$$

where b_f is the forecast profile, and H_f is the covariance matrix of forecast errors.

Another possibility is to use continuity along the suborbital track in the case of satellite measurements. Theoretically, we could treat all the observations made during one orbit, say, as one observation and retrieve the meteorological structure as a function of horizontal distance as well as height. Unfortunately, this is out of the question because of size of the matrices involved. However, we can use the statistics of horizontal temperature differences, and include the following pdf as a constraint:

$$-2 \log_e P(b) = (b - b_p)^T H_{\Delta}^{-1} (b - b_p)$$

where b_p is the previous retrieval and H_{Δ} is the covariance matrix of differences between profiles at the appropriate distance apart.

THE NONLINEAR CASE

The notation we have used so far can be generalized for the nonlinear problem. The vector b becomes a set of parameters describing the meteorological structure, not necessarily linearly. The observation I is now a known function $A(b)$ of the structure with the addition of experimental noise ϵ .

The most straightforward method of solution is to linearize the problem, and solve it with one of the linear methods described above, iterating until convergence is obtained. If the problem initially is nearly linear, a careful linearization may eliminate the need for iteration.

A generalization of the minimum variance can be stated as follows: Given

$$I = A(b) + \epsilon$$

find the function $D(I) = b^*$ such that

$$\sum (b - b^*)^2$$

is minimum for a given sample of profiles. This variance could be generalized for example to

$$(b - b^*)^T Q (b - b^*)$$

for some suitable matrix Q . Stated thus, the problem is difficult. The maximum probability method is more tractable.

We can write

$$-2 \log_e P(l|b) = (I - A(b))^T E^{-1} (I - A(b)) + \text{const}$$

and

$$-2 \log_e P(b) = (b - \bar{b})^T H^{-1} (b - \bar{b})$$

The solution that maximizes $P(b|l)$ can now be found by minimizing the explicit function of b

$$(I - A(b))^T E^{-1} (I - A(b)) + (b - \bar{b})^T H^{-1} (b - \bar{b}) \quad (11)$$

using a general numerical minimizing algorithm [e.g., *Powell*, 1964]. For particular forms of $A(b)$ it may be possible to construct more efficient minimizing routines, or even to solve Eq. (11) algebraically. If the solution is performed numerically it is advantageous to reduce as far as possible the number of parameters used to describe the structure. This can conveniently be done by using empirical orthogonal functions.

It is common in meteorological remote sounding for the observations to be linear in some quantities such as Planck function, but nonlinear in other quantities such as absorber distributions. It is possible to separate the linear and nonlinear parts of the problem in terms of the maximum probability method [*Rodgers*, 1970]. We can put

$$I = A(b) + \epsilon = K(n) \cdot l + \epsilon$$

where l is the linear part of b and n is the nonlinear part. The kernel $K(n)$ is a function of absorber distribution. Using Bayes theorem we can put

$$P(b|I) = P(l, n|I) = P(I|n, l) P(l|n) P(n) / P(I) \quad (12)$$

where the pdf are of the forms

$$\begin{aligned} -2 \log_e P(I|n, l) &= (I - K(n) \cdot l)^T \cdot E^{-1}(n) \cdot (I - K(n) \cdot l) \\ -2 \log_e P(l|n) &= (l - \bar{l}(n))^T \cdot H^{-1}(n) \cdot (l - \bar{l}(n)) \end{aligned} \quad (13)$$

where the covariance of l , H , and mean value of the linear part \bar{l} may depend on n . If we keep n fixed and solve for the most likely l , we obtain the same solution as before (7):

$$l^* - \bar{l}(n) = HK^T(KHK^T + E)^{-1}(I - \bar{l}(n)) \quad (14)$$

On substituting this and Eq. (13) back in (12) we obtain

$$-2 \log_e P(n, I|I) = [I - T(n)]^T \cdot X(n) \cdot [I - T(n)] - 2 \log_e P(n) + \text{const} \quad (15)$$

where

$$X(n) = (K(n)H(n)K^T(n) + E)^{-1}$$

This quantity can now be minimized numerically with respect to the nonlinear parameters, and the resulting value of n^* substituted in (14) to obtain the linear parameters l^* . The matrix X is the inverse of the covariance matrix of l for given n . Thus the quantity being maximized to find n^* is essentially $P(n|I)$.

COMMENTS

It is not possible in a review such as this to illustrate the practical use of all the retrieval methods discussed. The accuracy of retrieval depends as much on the particular radiances measured as on the retrieval method. It depends on the number, shape, and location of spectral intervals; the spatial resolution of the instrument; experimental accuracy; the precise nature of the atmospheric model; and many other things. Discussions of this aspect of the subject can be found in *Mateer* [1965], *Rodgers* [1970], *Strand and Westwater* [1968a,b], *Wark and Fleming* [1966], *Conrath* [1968, 1969], and *Turchin et al.* [1969].

DISCUSSION

Chahine: With the statistical method, clearly one can get the most probable profile. What about the least probable profile?

Rodgers: That is a slightly philosophical matter. Given an observation you want to find some solution. You have an infinite number. Which one do you want? The one I want is the one most likely to be right.

Chahine: As a physicist I am interested in the unusual. I would like to find that when it exists.

Rodgers: If a much more likely profile is just as consistent with the observations this is surely the answer you want.

Falcone: If you take the Bayesian technique, it reduces directly to Westwater-Strand's method, because you've used gaussian distributions. If you don't and you don't maximize with respect to B, you can develop this system into a Kallman filter system, which yields Smith's result, except that your iteration scheme gives you a weighting function which you get from the theory. You can get a mathematically good theory as Chahine wants, by looking at Tickenoff's method. He uses compact functions, and this is a good theory. Mathematically, it has its basis and it reduces to the Westwater-Strand method. This allows you to play the nonlinear game, the same as Staelin's method.

Rodgers: I think this is the most important point. You want to choose the method that you can generalize to the nonlinear case.

Fleming: The standard technique for solving nonlinear problems is to take the nonlinear equations and approximate them by a sequence of linear equations, which is what has been done in these two papers. Why does Chahine insist we solve directly a nonlinear problem?

Chahine: In the 15μ case that Dr. Rodgers is treating, the condition factor is of the order of 10^5 , which means that small noise or error in measurements will be amplified by this factor, making the solution obtained by this linearization process inaccurate.

Smith: That is the whole point of the error covariance matrix, which is to dampen the error.

Chahine: When you are using a damping factor you should look at how much information you are getting from your measurement and how much information you are getting from your damping function. All the significant information you are getting is coming from the latter.

Waters: To what extent does the accuracy of your solution depend on the statistics you use—for example, locally and one season versus globally and all seasons—and is there an optimum way of breaking the latter down?

Rodgers: We haven't been able to do that as yet. There is not an optimum way of breaking this down, but a lot of research is going into it.

**A REVIEW OF NONSTATISTICAL TECHNIQUES FOR THE
ESTIMATION OF VERTICAL ATMOSPHERIC STRUCTURE
FROM REMOTE INFRARED MEASUREMENTS**

B. J. Conrath and I. Revah*

Goddard Space Flight Center, Greenbelt, Maryland

N73-11500

ABSTRACT

Mathematical techniques used in atmospheric profile inversion are reviewed with attention limited to methods that do not require extensive information on the statistical properties of the profiles. The basic principles of the calculation of vertical profiles from noise contaminated infrared spectral measurements is examined. Three specific examples of methods of solution are given: the iterative minimum estimation method, a second linear iterative method, and the algorithm of Chahine. Satellite data are used in an empirical comparison of the methods. Several factors are considered, including the stability of solutions, convergence behavior of the iterative methods, and dependence of the solutions on the first guess, a factor that is particularly important in the treatment of data from other planets.

INTRODUCTION

The first explicit suggestion that temperature profiles in the terrestrial atmosphere could be obtained from measurements performed from an earth satellite was made by King [1956]. It was proposed that limb-scan measurements of thermally emitted atmospheric radiation be used for this purpose. Kaplan [1959] pointed out that measurements obtained in selected intervals within an infrared absorption band with a nadir-viewing instrument could be used for the same purpose. In the years following, a considerable amount of theoretical effort was devoted to the development of computational techniques for obtaining temperature profiles from the appropriate measurements.

Infrared measurements obtained from early meteorological satellites covered broad spectral bands and could not be used to obtain vertical profiles. In 1964 the first measurements suitable for temperature profile inversion were obtained from a balloon-borne spectrometer [Hilleary, *et al.*, 1965]; this work was followed by subsequent balloon flights of other instruments. However, it was not until the launching of the Nimbus 3 meteorological satellite on April 14, 1969, that suitable measurements were obtained on a global basis. That satellite carried two instruments that obtained infrared measurements at relatively high spectral resolution: the satellite infrared spectrometer (SIRS) [Wark and Hilleary, 1969] and the infrared interferometer spectrometer (IRIS) [Hanel and Conrath, 1969]. The Nimbus 4 satellite launched April 9, 1970, carried improved versions of SIRS and IRIS [Hanel and Conrath, 1970] as well as a selective copper radiometer (SCR). Thus, a large body of data now exists that can be used in, and has stimulated, the development of temperature profile retrieval techniques.

*NAS/NRC Resident Research Associate. Present address: Centre National d'Etudes des
Télécommunications, Paris, France.

A theoretical expression for the radiance $I(\nu)$ at wavenumber ν measured by a satellite-borne instrument can be obtained from a solution of the radiative transfer equation written in integral form

$$I(\nu) = B(\nu, T_s) \tau(\nu, p_s) - \int_0^{p_s} B[\nu, T(p)] \frac{\partial \tau(\nu, p)}{\partial p} dp \quad (1)$$

where $B(\nu, T)$ is the Planck function at wavenumber ν and temperature T . The atmospheric pressure p has been used here as the independent height variable, but any monotonic function of p can be used with $\ln p$ frequently being found convenient. The atmospheric transmittance from level p to the top of the atmosphere for a spectral interval centered at wavenumber ν is denoted by $\tau(\nu, p)$, and T_s and p_s refer to surface temperature and atmospheric pressure at the planetary surface, respectively. In deriving (1), it has been assumed that the atmosphere is nonscattering and in local thermodynamic equilibrium, and that the surface radiates as a black body.

The first term on the right-hand side of (1) represents the contribution from the planetary surface, while the second term is the atmospheric contribution. For sufficiently transparent spectral intervals, the boundary term must be specified in some manner, using measurements in the most transparent parts of the spectrum. It will be assumed that this can be done satisfactorily, and the boundary term in (1) will not be considered further here. Given measurements of $I(\nu)$, the basic problem then is to solve (1) to obtain an estimate of the vertical temperature profile $T(p)$.

To be able to solve for $T(p)$, it is necessary to know the atmospheric transmittance $\tau(\nu, p)$. Therefore, an absorption band for an atmospheric constituent of known abundance must be employed so that from a knowledge of the gaseous absorption coefficients, it is possible to calculate the atmospheric transmittance. Both CO_2 with absorption bands in the infrared and O_2 , which absorbs in the microwave region of the spectrum, satisfy this requirement. Inspection of (1) indicates that the vertical derivatives of the transmittance, or so-called "weighting functions," essentially define those portions of the atmosphere sampled by measurements at various points in the spectrum. Examples of weighting functions for 2.8 cm^{-1} wide spectral intervals in the $15 \mu\text{m}$ CO_2 band are shown in figure 1.

The techniques developed for the solution of (1) can be grouped roughly into two broad categories: statistical and nonstatistical. This discussion is confined to techniques that do not utilize extensive a priori statistical information about the atmospheric profiles.

The basic inversion problem is examined first, and three methods of solution typical of those commonly used today are discussed. Intercomparisons of the methods are made with data in the $15 \mu\text{m}$ CO_2 band obtained with the Nimbus 4 IRIS instrument.

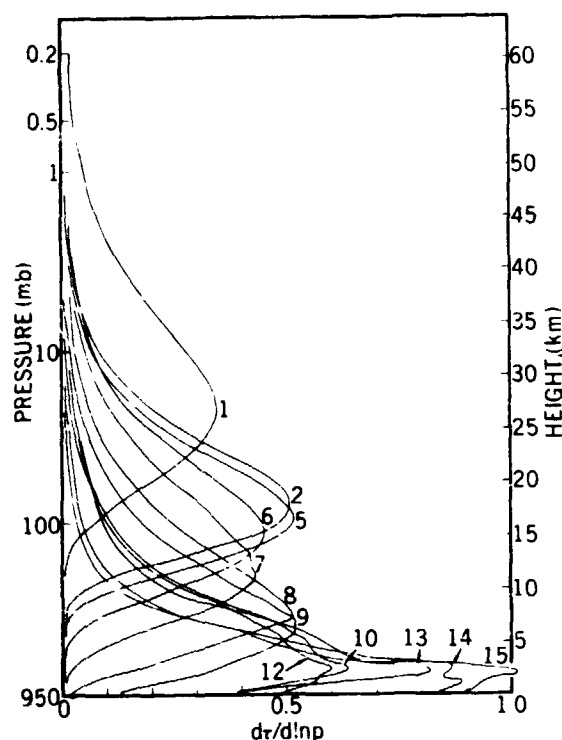


Figure 1. — Weighting functions for the $15 \mu\text{m}$ CO_2 absorption band. The absolute values of the derivatives of the transmittance with respect to the logarithm of the atmospheric pressure are shown for several spectral intervals 2.8 cm^{-1} wide.

METHODS OF SOLUTION

In general, measurements of $I(\nu)$ are available for some finite number of spectral intervals. If quantities pertaining to the spectral interval centered on wavenumber ν_i are identified with the subscript i , then (1) reduces to a set of m equations

$$I_i = - \int_0^{p_s} B_i [T(p)] \frac{\partial \tau_i(p)}{\partial p} dp \quad i = 1, 2, \dots, m \quad (2)$$

In the form given in (2), the equations to be solved are nonlinear in the unknown $T(p)$. Most methods of solution begin by linearizing in some way. (An exception to this approach is discussed later.) One form of linearization employs an expansion of $B[\nu, T(p)]$ about a reference profile $T^{(o)}(p)$ —that is

$$B_i [T(p)] \cong B_i [T^{(o)}(p)] + \frac{dB_i [T^{(o)}(p)]}{dT} [T(p) - T^{(o)}(p)] \quad (3)$$

Substitution of (3) into (2) results in a linearized set of equations of the form

$$\Delta I_i = - \int_0^{p_s} \frac{\partial \tau_i(p)}{\partial p} \frac{dB_i [T^{(o)}(p)]}{dT} \Delta T(p) dp \quad (4)$$

where $\Delta T(p) = T(p) - T^{(o)}(p)$ and $\Delta I_i = I_i - I_i^{(o)}$ with $I_i^{(o)}$ being the radiance calculated using $T^{(o)}(p)$. Approximating the integrals by numerical quadrature, (4) becomes

$$\Delta I_i = \sum_{j=1}^n K_{ij} \Delta T_j \quad i = 1, 2, \dots, m \quad (5)$$

where

$$K_{ij} = \frac{dB_i [T^{(o)}(p_i)]}{dT} \frac{\partial \tau_i(p_j)}{\partial p} w_j \quad (6)$$

with w_j an appropriate quadrature weight. Other forms of linearization have also been used [Wark and Fleming, 1966; Rodgers, 1966; 1970].

When linearization is employed, the problem is formally reduced to that of finding solutions to the linear set of equations (5). In the overdetermined case ($m > n$), a least squares solution can be written in vector form

$$\hat{\Delta T} = (K^T K)^{-1} K^T \Delta I \quad (7)$$

where $\hat{\Delta T}$ is the solution estimate and the superscript T denotes matrix transposition. In the atmospheric profile inversion problem, there will be generally many more quadrature points employed in (5) than measurements. In this underdetermined case ($m < n$), the family of solutions satisfying (5) can be written [Bjerhammar, 1951]

$$\Delta T = K^T(KK^T)^{-1} \Delta I + [K^T(KK^T)^{-1} K - I] b \quad (8)$$

where b is an arbitrary vector. With the additional constraint that $\Delta T \cdot \Delta T$ be a minimum, it can be shown that the resulting particular solution corresponds to $b = 0$ in (8), or

$$\hat{\Delta T} = K^T(KK^T)^{-1} \Delta I \quad (9)$$

which is the pseudoinverse of Penrose [Greville, 1959] for the underdetermined case.

In practice, solutions (7) and (9) are of little value for the atmospheric profile inverse problem. The presence of noise in the measured radiances renders the solutions physically meaningless. This is due to the high correlation among the measurements, and the resulting tendency toward redundancy of the equation set (5).

If an ensemble of measurements is considered, the covariance matrix for the measurements R can be written

$$R = \langle \Delta I \Delta I^T \rangle = KSK^T \quad (10)$$

where the angular brackets denote ensemble averaging, and S is the covariance matrix for the temperature profile,

$$S = \langle \Delta T \Delta T^T \rangle \quad (11)$$

The second form of R in (10) is obtained using (5). Now the correlation in the measurements is due both to the correlation of the atmospheric temperature among various atmospheric levels and the radiative transfer process itself—that is, the overlap of the weighting functions. The correlation due to the overlap of the weighting functions alone can be considered by assuming $S = \sigma_T^2 I$ where σ_T^2 is the variance in temperature assumed the same at each level. This results in

$$R = \sigma_T^2 KK^T \quad (12)$$

Thus, the matrix KK^T provides a measure of the correlation among the measurements due to the overlap of the weighting functions. It is possible to find a representation for the "measurement vector" ΔI in which its components are uncorrelated by diagonalizing KK^T . The resulting normalized eigenvectors u_i given by

$$KK^T u_i = \lambda_i u_i \quad i = 1, 2, \dots, m \quad (13)$$

provide a convenient basis set for representing ΔI in the measurement space and will be used in the subsequent analysis.

The effects of noise on (9) will now be considered. Assume the measured value $\tilde{\Delta I}$ consists of a "true" value ΔI , plus an error vector ϵ —that is

$$\tilde{\Delta I} = \Delta I + \epsilon \quad (14)$$

The noise-contaminated solution calculated from Eq. (9) is

$$\hat{\Delta T} = K^T (K K^T)^{-1} \tilde{\Delta I} \quad (15)$$

Now $\tilde{\Delta I}$ can be expanded in terms of u_i

$$\tilde{\Delta I} = \sum_{i=1}^m a_i u_i \quad a_i = u_i \cdot \tilde{\Delta I} \quad (16)$$

Substitution of Eq. (16) into Eq. (15) gives

$$\hat{\Delta T} = \sum_{i=1}^m \frac{a_i}{\lambda_i} K^T u_i \quad (17)$$

Thus, the solution can be written as a linear combination of the m vectors $K^T u_i$. These vectors form an orthogonal set but are not normalized, their lengths being $\sqrt{\lambda_i}$. Normalization yields the set of vectors

$$v_i = \frac{1}{\sqrt{\lambda_i}} K^T u_i \quad i = 1, 2, \dots, m \quad (18)$$

They are identical with the m eigenvectors of the matrix $K^T K$, which correspond to nonzero eigenvalues.

Using Eqs. (14) and (18), Eq. (17) can be rewritten in the form

$$\hat{\Delta T} = \sum_{i=1}^m \frac{u_i \cdot \Delta I + u_i \cdot \epsilon}{\sqrt{\lambda_i}} v_i \quad (19)$$

and the solution can be regarded as a linear combination of the vectors v_i in the "profile space." The vectors v_i were calculated for a set of 16 spectral intervals in the $15 \mu\text{m}$ CO_2 band. The first eight of these vectors, along with the corresponding eigenvalues, are shown in figure 2. It should be noted that there is a large dispersion in the range of eigenvalues corresponding to approximately 5 orders of magnitude with the more highly structured eigenvectors associated with smaller eigenvalues.

It is convenient to write $u_i \cdot \epsilon = |\epsilon| \cos \phi_i$ where $|\epsilon|$ is the magnitude of ϵ , and ϕ_i is the angle between ϵ and the i th basis vector u_i . Then Eq. (19) becomes

$$\hat{\Delta T} = \sum_{i=1}^m \left(\frac{u_i \cdot \Delta I}{\sqrt{\lambda_i}} + \frac{|\epsilon| \cos \phi_i}{\sqrt{\lambda_i}} \right) v_i \quad (20)$$

The first term of the coefficient of each v_i corresponds to the signal portion while the second term corresponds to the noise portion. Now if the true profile is relatively smooth, ΔI will be almost perpendicular to those u_i associated with small eigenvalues in the measurement space, and the first term in the corresponding coefficients in Eq. (20) will not be large even though $\sqrt{\lambda_i}$ appears in the denominator. However, since the noise vector will generally be random in direction in the measurement space, it can have nonvanishing components, in the direction of the u_i associated with small eigenvalues, that will be amplified by $1/\sqrt{\lambda_i}$. Thus, the coefficients of those v_i in the profile space that contain finer structure will be large, resulting in nonphysical, large-amplitude, fine-scale structure in the solutions.

Detailed analyses of the information content of noise contaminated measurements employed in obtaining vertical atmospheric profiles have been provided by *Mateer* [1965] and *Twomey* [1965]. Examples of the gross instabilities occurring when direct solutions are employed have been given in the literature [*Wark and Fleming*, 1966; *Conrath*, 1968].

To produce physically meaningful solutions, an inversion method must be able to suppress either explicitly or implicitly the contributions from those eigenvectors associated with the relatively small eigenvalues. Three computational methods typical of those currently being employed will now be discussed.

The first method to be considered is based on a form that can be derived as a limiting case of the statistical estimation

$$\hat{\Delta T} = SK^T(KSK^T + E)^{-1} \tilde{\Delta I} \quad (21)$$

obtained independently by *Rodgers* [1966] and by *Strand and Westwater* [1968 a,b]. The relation Eq. (21) was also given by *Foster* [1961] in a different context. Here $\hat{\Delta T} = \hat{T} - T^{(o)}$ where \hat{T} is the estimated profile, $T^{(o)}$ is the mean of a representative ensemble of profiles, S is the two-level profile covariance matrix obtained from the same ensemble, $\tilde{\Delta I} = \tilde{I} - I^{(o)}$ where $I^{(o)}$ are the radiances calculated using $T^{(o)}$, and E is the instrumental noise covariance matrix. The limiting form of Eq. (21) of interest here is obtained by assuming $S = \sigma_T^2 1$ and $E = \sigma_\epsilon^2 1$;

$$\hat{\Delta T} = K^T(KK^T + \gamma 1)^{-1} \tilde{\Delta I} \quad (22)$$

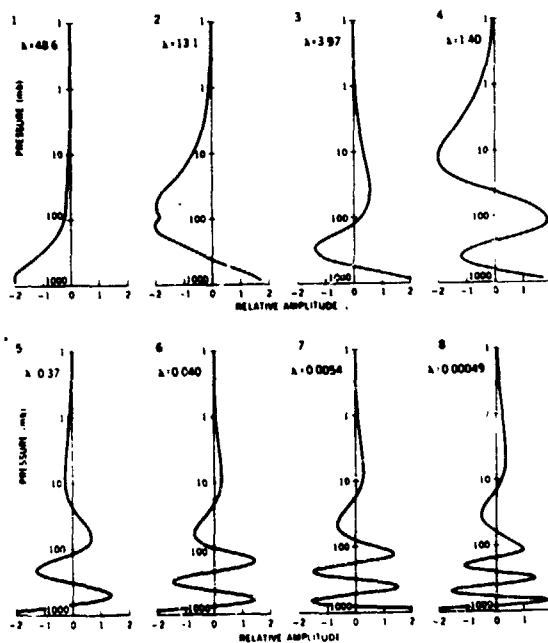


Figure 2. — Basis vectors employed in the representation of solutions in the profile space. The associated eigenvalue is shown in each case.

where $\gamma = \sigma_e^2 / \sigma_T^2$ is the ratio of noise variance to signal variance. This is called the *minimum information*, or *maximum entropy*, solution by Foster [1961]. A similar algebraic form was derived by Twomey [1963] from a nonstatistical point of view based on the work of Phillips [1962]. A similar form can also be obtained from the method of regularization [Glasko and Timofeyev, 1968].

When $T^{(0)}$ and γ are not precisely known from statistics, Eq. (22) can be used as the basis for an iterative form, treating $T^{(0)}$ as a first-guess profile. The resulting iterative relation is

$$T^{(\ell+1)} = T^{(\ell)} + K^T (KK^T + \gamma I)^{-1} [\tilde{T} - I^{(\ell)}] \quad (23)$$

where the parenthetical superscripts refer to the iteration order. Iteration can continue until the residuals $\tilde{T} - I^{(\ell)}$ reach the instrumental noise level.

The behavior of this solution can be examined by considering the representation of the residuals on the ℓ th iterative step in terms of the eigenvectors u_i

$$\tilde{T} - I^{(\ell)} = \sum_{i=1}^m a_i^{(\ell)} u_i \quad a_i^{(\ell)} = [\tilde{T} - I^{(\ell)}] \cdot u_i \quad (24)$$

From Eqs. (24) and (23), a recursion relation for the coefficients $a_i^{(\ell)}$ can be derived in the form

$$a_i^{(\ell+1)} = \frac{\gamma}{\lambda_i + \gamma} a_i^{(\ell)} \quad (25)$$

from which the relationship of the coefficients on the ℓ th iteration to their initial values can be obtained

$$a_i^{(\ell)} = \left(\frac{\gamma}{\lambda_i + \gamma} \right)^\ell a_i^{(0)} \quad (26)$$

The mean square residual on the ℓ th iteration, which can be employed as a measure of convergence, can be written

$$\frac{1}{m} (\tilde{T} - I^{(\ell)}) \cdot (\tilde{T} - I^{(\ell)}) = \frac{1}{m} \sum_{i=1}^m \left(\frac{\gamma}{\lambda_i + \gamma} \right)^{2\ell} (a_i^{(0)})^2 \quad (27)$$

Note that for $\gamma > 0$ the terms in Eq. (27) will individually approach zero in the limit $\ell \rightarrow \infty$. The rate of the reduction of each term with increasing ℓ depends on the relationship of γ to λ_i . Terms for which $\gamma \ll \lambda_i$ will decrease very rapidly while those for which $\gamma \gg \lambda_i$ will decrease slowly. Thus, by choosing an appropriate value of γ , the portion of the residuals associated with the larger eigenvalues and valid information can be reduced in one or a few iterations, while that portion associated with small eigenvalues and noise will remain essentially unreduced. By truncating the iteration at the proper mean square residual level, the noise contaminated modes associated with small λ_i values can be suppressed.

The solution on the ℓ th iteration in terms of the set of basis vectors v_i in the profile space can be written

$$T^{(\ell)} = T^{(0)} + \sum_{i=1}^m \left[1 - \left(\frac{\gamma}{\lambda_i + \gamma} \right)^\ell \right] \frac{a_i^{(0)}}{\sqrt{\lambda_i}} v_i \quad (28)$$

This expression can be obtained from Eqs. (23) and (24) by straightforward algebraic manipulation. It can be seen that the coefficients of those v_i for which $\gamma \gg \lambda_i$ will increase very slowly, and these terms will not contribute to the solution until many iterations have been performed. In the limit $\ell \rightarrow \infty$, the factor in square brackets approaches unity for all $\gamma > 0$ and Eq. (28) reduces to an expression equivalent to Eq. (19), indicating that the iterative algorithm Eq. (23) converges to the Penrose pseudoinverse. In practice, the iteration is stopped when the residuals reach the instrumental noise level.

Another type of iterative algorithm can be obtained in the following way. Multiplying both sides of Eq. (5) by K^T gives

$$K^T \Delta I = K^T K \Delta T \quad (26)$$

Now consider $K^T K$ written as the sum of a purely diagonal matrix G and a remainder R ; that is

$$K^T K = G + R \quad (27)$$

Once G is specified, R is determined by Eq. (27). Combining Eq. (27) with Eq. (26) and rearranging gives

$$G \Delta T = K^T \Delta I - R \Delta T \quad (28)$$

This expression can serve as the basis of an iterative algorithm

$$\Delta T^{(\ell+1)} = G^{-1} K^T \tilde{\Delta I} - G^{-1} R \Delta T^{(\ell)} \quad (29)$$

where $\Delta T^{(\ell+1)} = T^{(\ell+1)} - T^{(0)}$ and $\Delta T^{(\ell)} = T^{(\ell)} - T^{(0)}$. Using Eq. (27) to eliminate R finally results in

$$T^{(\ell+1)} = T^{(\ell)} + G^{-1} K^T (\tilde{I} - I^{(\ell)}) \quad (30)$$

The choice of G determines the convergence behavior of Eq. (30). Examples of solutions are given in the following section with the n diagonal elements of G set equal to

$$\sum_{k=1}^m \sum_{j=1}^m K_{jk} K_{ji} \quad i = 1, 2, \dots, n.$$

A third example of a nonstatistical solution is that suggested by *Chahine* [1968]. It is a relaxation technique that does not require linearization of the radiative transfer equation, but uses Eq. (2) directly. In this approach, the radiance from the i th spectral interval is paired with the temperature at some level p_i . The algorithm employed is

$$B_i \left[T^{(\ell+1)}(p_i) \right] = \frac{\tilde{I}_i}{I_i^{(\ell)}} B_i \left[T^{(\ell)}(p_i) \right] \quad i = 1, 2, \dots, m \quad (31)$$

The superscripts refer to the iteration number, and $I_i^{(\ell)}$ is the radiance calculated using the estimate for the temperature profile on the ℓ th iteration. A first guess is made at a temperature profile to begin the iteration. In order to calculate $I_i^{(\ell)}$ at each iterative step, the temperatures $T^{(\ell)}(p_i)$ at the m atmospheric levels must be interpolated to intermediate levels. The form of interpolation used, along with the quadrature method employed, essentially defines the representation for the temperature profile.

Iteration is carried out using Eq. (31) until some condition on the residuals (differences between measured and calculated radiances) is met. For example, iteration may be stopped when the rms residual approaches some value determined from a knowledge of the rms noise level of the measurements. The convergence properties of Eq. (31) are rather obscure, and the degree of success of the method must be judged empirically. There is obviously no unique choice of atmospheric level p_i for a given spectral interval, but a logical choice would appear to be a pressure level near the peak of the weighting function for that interval.

A modification of Chahine's algorithm has been given by *Smith* [1970]. This approach employs an iterative equation

$$B_i^{(\ell+1)}(p) = B_i^{(\ell)} + \left[\hat{I}_i - I_i^{(\ell)} \right] \quad i = 1, 2, \dots, m \quad (32)$$

By solving the Planck functions obtained with this equation for temperature, m different temperature profiles $\hat{T}_i^{(\ell+1)}(p)$ ($i = 1, 2, \dots, m$) are obtained. The solution is then defined as the weighted mean

$$T^{(\ell+1)}(p) = \frac{\sum_{i=1}^m K_i(p) \hat{T}_i^{(\ell+1)}(p)}{\sum_{i=1}^m K_i(p)} \quad (33)$$

A first guess at a temperature profile is made, and Eqs. (32) and (33) are iterated until some criterion on the radiance residuals is satisfied. This method does not require an arbitrary pairing of atmospheric levels with spectral intervals nor does it require an explicit interpolation between levels since a temperature can be calculated at each level using Eq. (33).

EMPIRICAL COMPARISONS OF METHODS

The behavior of three of the methods discussed above are now examined and compared, using applications to actual satellite data. These methods include the minimum information estimation Eq. (23), Chahine's algorithm Eq. (31), and the iterative algorithm Eq. (30).

The data were obtained with the Nimbus 4 IRIS experiment [Hanel and Conrath, 1970]. The instrument has been described in detail elsewhere [Hanel et al., 1971]. The spectral range extends from 400 cm^{-1} to 1500 cm^{-1} , and the spectral resolution is equivalent to about 2.8 cm^{-1} . Selected spectral intervals in the $15\text{ }\mu\text{m}$ (667 cm^{-1}) CO_2 band are used in the present study. The noise equivalent radiance in this spectral region is approximately $0.5\text{ erg/cm}^2\text{ sec sr cm}^{-1}$ (1% of the signal). The quality of the solutions is established by comparisons with in situ measurements obtained with radiosondes.

Examples of Retrieved Profiles

The first examples to be considered are solutions obtained with each of the three techniques, compared with a radiosonde profile. Sixteen spectral intervals were employed; weighting functions for several of the intervals are shown in figure 1. Figure 3 shows the retrievals using the minimum information method, the iterative estimation method Eq. (30), and the Chahine algorithm. In each case, the same first guess, a climatological mean profile, was used.

The minimum information solution and the iterative method are found to give almost identical results in this case: they are indistinguishable on the scale of figure 3. The minimum information method was iterated twice with $\gamma = 0.03(^{\circ}\text{K})^2/(\text{erg/cm}^2\text{ sec sr cm}^{-1})^2$ to reduce the residuals to the instrumental noise level. The temperature dependence of K was taken into account on each iteration. The solution shown in figure 3, using the iterative method Eq. (30), was obtained with 24 iterations, but is essentially no different from that obtained with only five iterations.

In obtaining the retrieval shown in figure 3 with the Chahine algorithm, linear interpolation was employed between the 16 levels for which temperatures could be computed. The appearance of spurious fine structure in the solution is a problem encountered when an attempt is made to use a large number of spectral intervals, as in this case. This structure is essentially the onset of instability and grows in amplitude as the number of iterations is increased. The use of 16 spectral intervals results in a large amount of redundancy with weighting function peaks separated only slightly in many cases. This makes the selection of levels for calculating temperatures using Eq. (31) a rather difficult task. With so many spectral intervals, the choice of criteria for truncating the iteration becomes critical. In studies with the Chahine method the iteration was stopped whenever the rms residual dropped below $0.8\text{ erg/cm}^2\text{ sec sr cm}^{-1}$ or changed by less than $0.02\text{ erg/cm}^2\text{ sec sr cm}^{-1}$ from the previous iteration. The instability problem disappears when the number of spectral intervals is reduced, and the use of 7 intervals produces better results than 16.

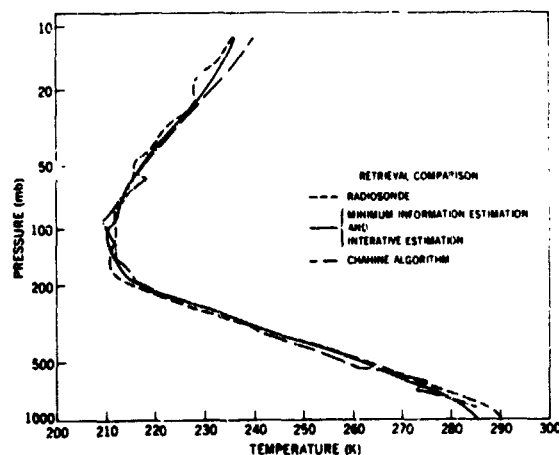


Figure 3. - Comparison of retrievals using three methods. The solutions obtained with the minimum information estimation and the iterative method are indistinguishable on this scale.

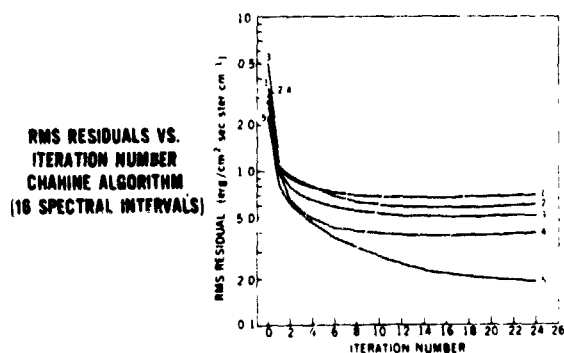


Figure 4.— Convergence behavior of the Chahine algorithm. The five examples shown were selected from a larger sample to show the range of behavior.

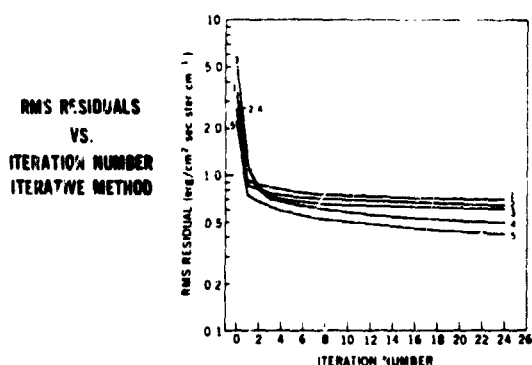


Figure 5. Convergence behavior of the iterative method. The same five cases employed in figure 4 are shown.

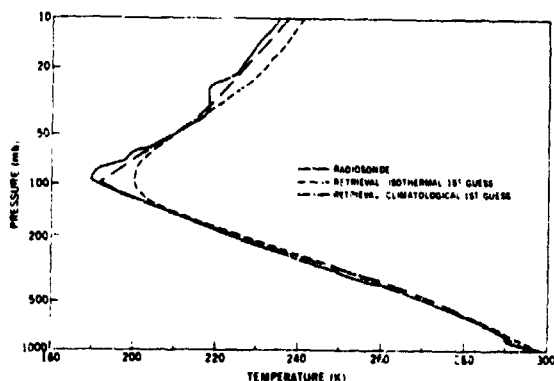


Figure 6.— Minimum information estimation retrievals illustrating dependence of solutions on the first guess. The isothermal first guess was 250° K.

Convergence Behavior

It was found empirically that iterating the minimum information estimation more than twice did not appreciably improve the solutions. Hence, this method can be done with considerable computational speed.

The convergence behavior of the Chahine algorithm is shown in figure 4 for five typical cases. The rms residual is plotted versus iteration number. The residual for the 0th iteration represents that obtained with the first-guess temperature profile. In general, all the curves are characterized by a large decrease in the residual on the first iteration, followed by a region of less rapidly decreasing residual until the curves become nearly flat. In some cases the residuals actually increase again after going through rather broad minima. There is considerable dispersion in the asymptotic behavior of the curves, which appears to depend on the nature of the profile, the first guess, and the behavior of the noise in each case.

The best results are obtained when the iteration is stopped at some point where the residuals are beginning to decrease less rapidly but have not yet reached the flat part of the curve. However, the optimum point for stopping the iteration appears to vary from case to case, making it difficult to establish a general set of convergence criteria. The quality of a solution is found to decrease rapidly as the flat part of the curve is approached.

The change in rms residual with iteration number for the iterative method is shown in figure 5 for the same five cases. The general shape of the curves is qualitatively similar to those obtained with the Chahine algorithm but with a somewhat more abrupt transition to the flat portion of the curve. The solutions were found not to change significantly after iteration 4 or 5, and the quality remained good out through iteration 24, which was as far as the calculations were carried. Thus, the criteria for stopping the iteration are not critical with this method.

Dependence of Solutions on the First Guess

Since all three solutions require a first-guess temperature profile, it is of interest to investigate the sensitivity of the final solutions to the first guess. For this purpose, a case was chosen in the tropics where the temperature profile has a very sharp minimum (tropopause). This represents essentially fine structure and as such is generally difficult to retrieve accurately.

Solutions were obtained with each of the three methods using two extremes in first guess: (1) a climatological mean temperature profile having nearly the same shape as the actual profile, and (2) an isothermal profile at 250° K. Figure 6 shows the results of this test for the minimum information estimation Eq. (23), with the iterative method Eq. (30) and Chahine's algorithm Eq. (31) giving essentially similar results.

Good results were obtained with all three methods using the climatological first guess. With the isothermal first guess, the gross structure of the profile is retrieved. The essentially linear part of the profile in the lower atmosphere (troposphere) is fit quite accurately. However, the solutions are about 10° K too warm at the tropopause, and the sharp structure is not retrieved.

SUMMARY

The mathematical techniques employed in obtaining atmospheric profiles have undergone extensive evolution since the early attempts at direct matrix inversion which resulted in highly unstable solutions. Various methods have been developed for obtaining stable solutions which are physically meaningful. These solutions generally employ constraints, either implicitly or explicitly, that may involve the use of extensive information on the statistical properties of the profile or may use only a knowledge of the general properties of the profile (for example—its smoothness). Both general approaches have recently been employed in the reduction of infrared spectral data acquired from satellites [Smith *et al.*, 1970; Wark, 1970; Conrath *et al.*, 1970]. The present review has been limited to techniques that do not use extensive statistical information.

Constrained forms of matrix inversion have been developed through the work of Phillips and Twomey, and more recently the so-called "minimum information estimation," closely related to the solutions given by Twomey, has been used. The minimum information solution can be viewed as either a limiting form of a statistical estimation or a constrained form of pseudoinversion for an underdetermined system. Several iterative techniques have been developed, including that of Chahine, which uses ratios of measured to calculated radiances, and other methods using differences between measured and calculated radiances.

Three methods typical of those currently in use were described and compared using actual satellite data. The performance of the methods also was compared against in situ measurements obtained with radiosondes. It should be emphasized that the empirical results obtained here pertain to the 15 μm CO₂ band and could conceivably be altered if other spectral bands were considered.

When large quantities of data are to be reduced for the terrestrial atmosphere on a global basis, computational speed is of utmost importance. The minimum information estimation is rapid, and therefore attractive for this application, because it requires few iterations. Increasing the number of spectral intervals increases the computational time required for all methods, and this factor must be weighed against any possible gain through redundancy.

Study of the convergence behavior of the Chahine algorithm indicates that the choice of criteria for truncating the iteration is very important. For the iterative method Eq. (30), the results are insensitive to the number of iterations within a rather broad range. A study of the dependence of the methods on the first guess indicates that only the gross features of the profile can be retrieved if a poor first guess is employed, such as an isothermal atmosphere. This aspect is of particular interest when the application of the techniques to spectra from other planets is considered. An example of such an application is the retrieval of profiles for the Martian atmosphere from spectra to be obtained with the IRIS instruments flown on the 1971 Mariner missions [Hanel *et al.*, 1970].

In this review of the mathematics of atmospheric profile inversion, several factors in reducing data for the earth's atmosphere on a global basis have not been discussed. Chief among these is the problem of a partially cloud-filled instrumental field of view. In this situation, the radiances must be corrected so the cloud effects are removed before the techniques described can be applied. The correction is accomplished using information from other spectral intervals or information derived from spatial scanning [Smith, 1967; Smith *et al.*, 1970; Rodgers, 1970].

While only the retrieval of temperature profiles has been considered in detail here, the same mathematical techniques can be employed to retrieve the vertical profiles of optically active gases. Once the temperature profile has been obtained from measurements in the absorption band of a gas whose abundance is known, measurements in absorption bands of other gases such as water vapor and ozone yield information on those gases. Adaptation of techniques, such as those described here, to the problem of retrieving information on atmospheric constituents has been described in the literature [Conrath, 1969; Prabhakara, 1969; Smith, 1970].

ACKNOWLEDGMENTS

The authors wish to thank Dr. R. A. Hanel, Principal Investigator of the IRIS Experiment, and Mr. V. G. Kunde, Dr. C. Prabhakara, and Dr. V. V. Salomonson for their cooperation during the course of this study.

DISCUSSION

Strand: I'd like to ask about the strong dependence of the initial guess. Why do you lose all this information and never get it back when you start with a poor initial guess?

Conrath: The structure, like the tropical tropopause, is not contained in your measurements. You need an initial guess containing this.

Waters: In applying this technique to atmospheres other than that of the earth, how well do you know the kernel in the integrand?

Conrath: You are on shakier ground. First, you don't know the mixing ratio precisely. This upsets the relationship between the optical depth and pressure level and introduces profile distortions. Also with low pressures and high mixing ratios (Mars) we are extrapolating quite a bit with our theoretical representations of the absorption coefficients. This brings in other uncertainties, too.

Falcone: If you use the Bayesian technique and you put in your a priori knowledge, then you can look at your variance to see how it is changing. You will get better knowledge that way.

Conrath: You essentially have no real a priori knowledge. All you have is the variance of your measurements.

Unidentified speaker: On your curves you presented of the rms radiance residuals versus iteration number, have you looked at what your rms temperature errors do?

Conrath: Yes. For the Chahine type relaxation algorithm if you keep iterating out on the flat part of the curve the temperature errors go through a minimum and then go up. For the other method it doesn't make a difference and you get a very flat minimum.

McClatchey: From what you have shown it appears that the problem is solved. All of the methods gave very good answers. Does that mean we can get the temperature information we want?

Conrath: That is a very strong statement and I am reluctant to agree. The subject of cloudy atmospheres is still a big problem.

Chahine: I would like to comment on the degree of vertical resolution we can get from a given set of measurements. The frequencies used to sound the tropopause are very broad and because the tropopause temperature is lower than above or below it, a large amount of outgoing radiance comes from the region above and below the tropopause. Therefore, the inverse problem will not reconstruct the tropopause. Methods should be developed to measure outgoing radiance with a high degree of accuracy and high spectral resolution.

McClatchey: I have generated from a series of real atmospheric temperature soundings radiance distributions for both the 15μ and 4μ cases for every two wavenumbers in each band and using a set of transmission functions which I would be glad to provide. I would like someone to do the inversion and I'll keep the numbers locked up. This should be a very good and interesting test of various techniques.

STATISTICAL VERSUS NONSTATISTICAL TEMPERATURE INVERSION METHODS

W. L. Smith and H. E. Fleming

National Environmental Satellite Service
National Oceanic and Atmospheric Administration
Hillcrest Heights, MD

N73-11589

ABSTRACT

Vertical temperature profiles are derived from radiation measurements by inverting the integral equation of radiative transfer. Because of the nonuniqueness of the solution, the particular temperature profile obtained depends on the numerical inversion technique used (that is, matrix inversion or iteration) and the type of auxiliary information incorporated in the solution (that is, statistical or initial data). The choice of an inversion algorithm depends on many factors, including the speed and size of computer, the availability of representative statistics, and the accuracy of initial data. This paper presents results of a numerical study comparing two contrasting inversion methods: the statistical-matrix inversion method and the nonstatistical-iterative method. These were found to be the most applicable to the problem of determining atmospheric temperature profiles. Tradeoffs between the two methods are discussed.

INTRODUCTION

The determination of vertical temperature profiles from multispectral satellite measurements of the earth's radiance, in an absorption band of a known gaseous constituent, requires the inversion of the integral equation of radiative transfer

$$I(\nu) = B[\nu, T(x_0)] \tau(\nu, x_0) - \int_0^{x_0} B[\nu, T(x)] \frac{d\tau(\nu, x)}{dx} dx \quad (1)$$

The spectral radiance measured from space is given by $I(\nu)$, $B(\nu, T)$ is the Planck radiance at frequency ν and temperature T , x is any single-valued function of pressure p , and $\tau(\nu, x)$ is the fractional transmittance of the atmosphere above the level x for radiation at frequency ν . In this discussion $x(p) = a [\ln(100p)]^2 + 1$ where $a = 99/[\ln(100,000)]^2$. The zero subscript denotes surface boundary values. The boundary term is determined from atmospheric "window" measurements. The solution $T(x)$ is obtained from the Planck radiance $B[\nu, T(x)]$ assuming the kernel of $d\tau/dx$ of Eq. (1) is known. Figure 1 shows the nine components of the kernel (or Planck radiance "weighting functions") for the spectral intervals of a future operational vertical sounder. The Planck radiance is given by

$$B(\nu, T) = c_1 \nu^3 / [\exp(c_2 \nu / T) - 1] \quad (2)$$

where c_1 and c_2 are constants of the Planck function.

Fleming and Smith [1971] give the details of various numerical algorithms for solving Eq. (1) for $T(x)$. Briefly, the solution of Eq. (1) requires measurements $I(\nu)$ to a certain degree of accuracy at a finite number of frequencies and knowledge of $\tau(\nu, x)$. Because of the conservative nature of the atmosphere, one begins with an initial profile for $T(x)$ —a climatological mean or a forecast profile—that reduces the problem to solving for the temperature deviation rather than for its entire magnitude. The Planck function is then linearized with respect to frequency about some intermediate reference frequency.

The resulting matrix-vectorial approximation of the radiative transfer equation is

$$r = A b + \epsilon \quad (3)$$

where r is a column vector of differences of observed radiances from radiances given by the initial data; A is an $M \times N$ matrix whose elements are

$$a_{ij} = -a_{ij} \frac{d\tau(\nu_i, x_j)}{dx} \quad \begin{array}{l} i = 1, \dots, M \\ j = 1, \dots, N \end{array} \quad (4)$$

where M is the number of measurement frequencies and N is the number of pressure levels. The elements of the column vector b are

$$b_j = B[\nu_0, T_j] - B[\nu_0, \bar{T}_j] \quad (5)$$

and the observational errors are represented by the row vector ϵ . The Planck function linearization factors in Eq. (4) are given by

$$\alpha_{ij} = \begin{cases} \frac{B(\nu_i, T_j) - B(\nu_i, \bar{T}_j)}{b(x_j)} & \text{for } T_j \neq \bar{T}_j \\ \frac{dB(\nu_i, T_j)/dT_j}{dB(\nu_0, T_j)/dT_j} & \text{for } T_j = \bar{T}_j \end{cases} \quad (6)$$

where ν_0 is the reference frequency and \bar{T}_j is the initial temperature at the j th pressure level. Since T_j in Eq. (6) is unknown, \bar{T}_j must be used initially, but on successive iterations the latest estimate of T_j can be used in estimate α_{ij} . The inverse problem is now reduced to solving Eq. (3) for b since the temperature profile T can be obtained from the unique inverse of the Planck function Eq. (2).

Equation (3) cannot be solved directly for b in that it is unstable because of the near singularity of the matrix A . The near singularity is due to the lack of vertical independence of the weighting functions (fig. 1). Numerous indirect methods of solving Eq. (3) have been presented

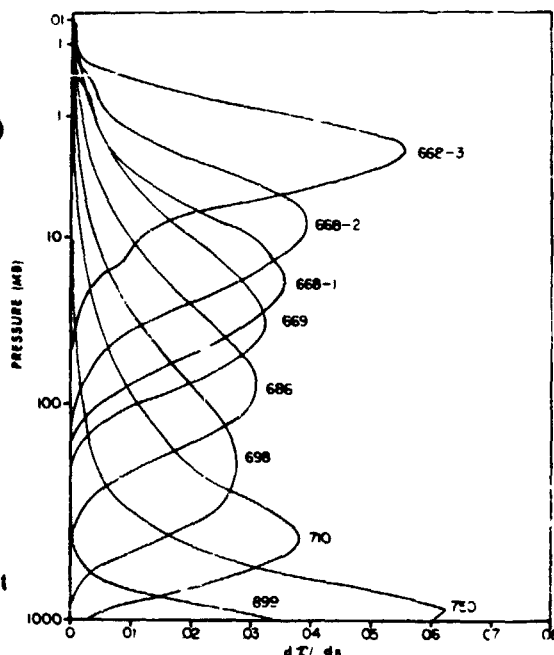


Figure 1.— Derivative of transmittance (weighting functions) versus pressure for the spectral wave numbers indicated.

in the literature. However, since the solution of Eq. (3) is nonunique, the solution one obtains depends on the method used.

METHODS

Two very different methods of solving Eq. (3) for \mathbf{b} are given in *Fleming and Smith* [1971]: the statistical matrix inversion method and the nonstatistical iterative method. The statistical method yields the solution

$$\mathbf{b} = \mathbf{C} \mathbf{r} \quad (7)$$

in which the coefficient matrix

$$\mathbf{C} = \mathbf{S}_b \mathbf{A}^T [\mathbf{A} \mathbf{S}_b \mathbf{A}^T + \mathbf{S}_\epsilon]^{-1} \quad (8)$$

where \mathbf{S}_b and \mathbf{S}_ϵ are covariance matrices of \mathbf{b} and ϵ . The superscripts T and -1 denote the matrix transpose and matrix inverse, respectively. In Eq. (7) the elements of \mathbf{b} are the deviations of the actual temperature from the mean temperature values obtained from the statistical sample of soundings used to estimate \mathbf{S}_b , and \mathbf{r} is the vector of deviations of the measured radiances from the radiances corresponding to the mean temperature profile.

The general iterative method utilizes the solution

$$\mathbf{b}(n+1) = \mathbf{C} \mathbf{r}(n) \quad (9)$$

where the parenthetical superscripts (n) represent the n th step in the iterative procedure. The coefficient matrix

$$\mathbf{C} = \mathbf{D} \mathbf{A}^T \quad (10)$$

where \mathbf{D} is a diagonal matrix with diagonal elements

$$d_{jj} = 1 / \left(\sum_{i=1}^M a_{ij} \right) \quad j = 1, \dots, N$$

In Eq. (9) \mathbf{b} and \mathbf{r} are deviations of atmospheric temperature and measured radiances as in the previous paragraph, but now the deviations are from the values calculated in the previous step of the iteration.

Basic differences in the two methods of solution are readily apparent from a comparison of the coefficient matrices given in Eqs. (8) and (10). The iterative method avoids the calculation of an inverse matrix, but at the expense of iteration. On the other hand, the statistical method utilizes statistical covariance matrices and an inverse matrix to achieve a one-step solution for the departure of atmospheric temperature from a fixed mean profile. The iterative method is free to use an arbitrary initial profile and attempts to successively improve the solution through iteration. Also, this method is easier to implement than the statistical method since a priori statistics are not required. However, the iterative method is more time consuming than the statistical method due to the number of iterations required.

RESULTS

The results presented here were obtained from calculations previously described by *Fleming and Smith*, [1971] in their numerical comparison of the two inversion methods. The procedure was essentially as follows. Radiances were calculated by Eq. (1) from a diverse sample of 109 temperature profiles between 15°N and 90°N latitude. Random noise, distributed normally with a standard deviation of 0.2 erg/(cm² sec sr cm⁻¹), was added to the radiances to simulate real measurements. The 109 profiles were then retrieved from the perturbed radiances by the statistical and the general iterative methods, Eqs. (7) and (9).

Covariance matrices S_b and their corresponding mean profiles \bar{T} were determined for four latitudinal zones—15-30°N, 30-45°N, 45-60°N, 60-90°N—in an independent study that did not utilize any of the 109 profiles used in the present study. The mean profiles used in the statistical method were also used as initial profiles in the iterative method for consistency. The noise covariance matrix S_e was simplified to the diagonal form $\sigma_e^2 I$ with $\sigma_e = 0.2$. The convergence criteria of the iterative method was based on $\sigma_e = 0.2$ as an upper bound.

Figure 2 shows the mean absolute temperature error associated with the two methods as a function of pressure. The accuracy of both methods is relatively poor in the layers of the atmosphere where abrupt changes in lapse rate exist; for instance, below 700 mb where surface inversions occur, within the 300 to 70 mb layer where the tropopause prevails, and in the upper stratosphere where the stratopause temperature reversal exists. In the layers where constant lapse rates of temperature generally prevail the accuracies of both methods is close to 1°C even though the simulated radiances were contaminated with realistic noise. The general saw-toothed nature of the curves is related to the fact that individual members of the set of 109 soundings contain sharp temperature inversions that occur randomly in layers having thicknesses below the resolution provided by the weighting functions.

Figure 3 shows the average difference of the absolute errors of the two methods for four latitude belts within the 15-90°N region. The error difference was computed by subtracting the average absolute error of the statistical solution from the average absolute error of the nonstatistical method for each latitude region. Consequently, layers of positive difference are layers in which the statistical solution is superior to the nonstatistical solution.

In viewing the difference curves for the four latitude belts, one sees that the only vertical layer possessing a consistently one-sided difference is the tropopause layer (300 to 70 mb). In this region of the atmosphere the average accuracy of the statistical solution exceeds

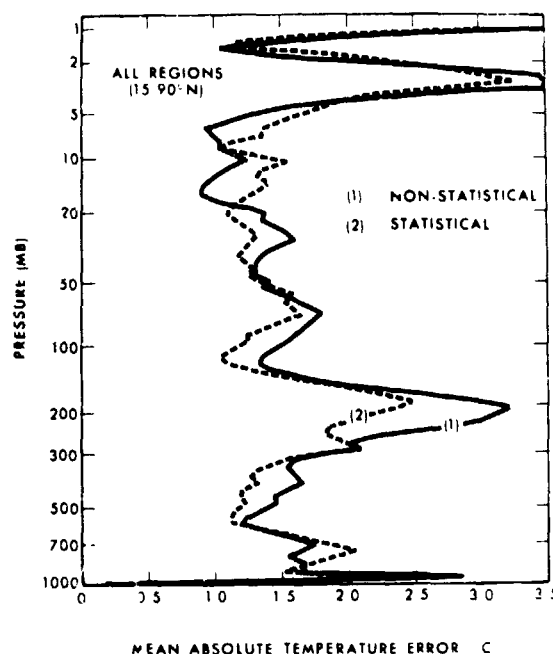


Figure 2.— Mean absolute temperature errors as a function of pressure for 15-90°N. Solid line and dashed line correspond to the two solutions indicated by the numbers.

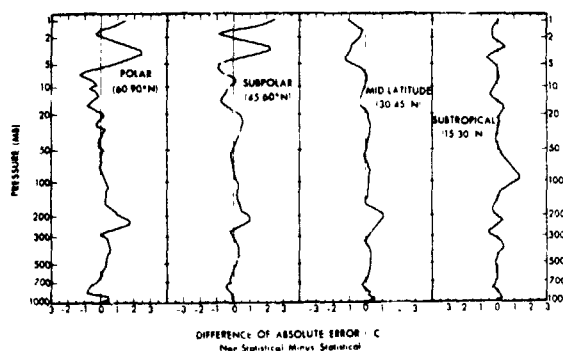


Figure 3.— Difference of absolute error between nonstatistical and statistical methods for polar, subpolar, midlatitude and subtropical areas.

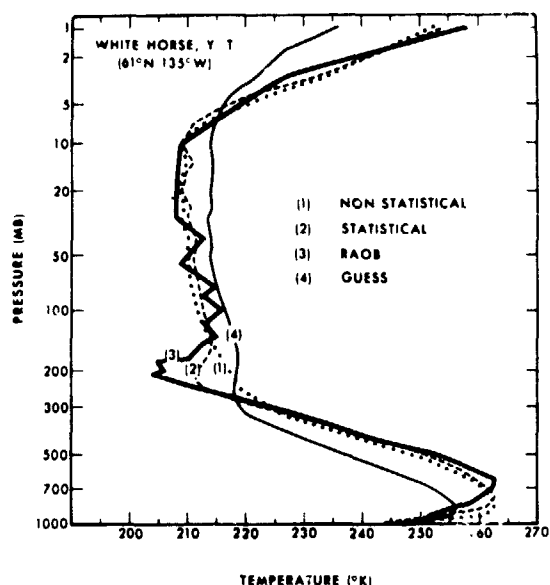


Figure 4.— Temperature profiles as a function of pressure for White Horse, Y. T. The RAOB, guessed profile, and the two solutions are identified by the numbers.

profiles in which the largest positive differences and the largest negative differences, respectively, of figure 3 occurred. The actual temperature profile (denoted RAOB for radiosonde observation) shown in figure 4 is characterized by a very sharp and cold tropopause. The tropopause layer is less than 100 mb thick and apparently below the resolution of the radiance observations since the nonstatistical solution failed to exhibit this tropopause feature. However, the

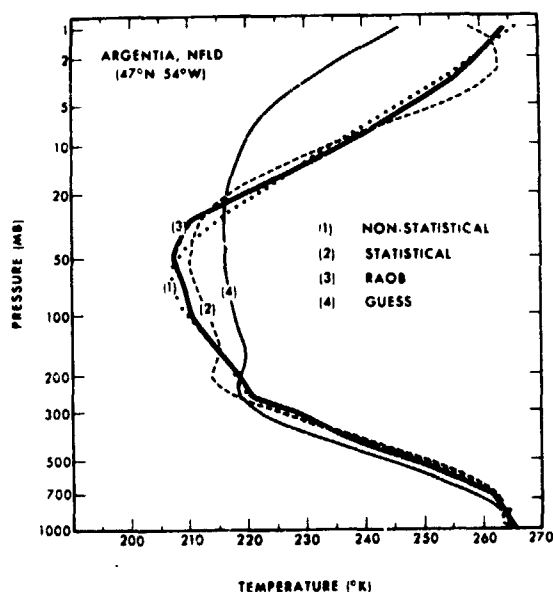


Figure 5.— Temperature profiles as a function of pressure for Argentina, Nfld. The RAOB, guessed profile, and the two solutions are identified by the numbers.

that of the nonstatistical solution by as much as 1°C . (The upward progression of the tropopause toward lower pressures with decreasing latitude is quite evident.) Statistics are expected to improve the accuracy in this layer because the small-scale tropopause features (its height and change in temperature lapse rate) are known to be highly correlated with broader scale characteristics, such as average lapse rate and vertical mean temperature, of the troposphere and stratosphere. It is somewhat surprising that the statistics seem to play an insignificant role within other regions of the atmosphere, except in the upper polar stratosphere. This indicates that smaller scale vertical features outside the tropopause layer are not well correlated with features that can be resolved by the radiances. In fact, there are regions—for example the layers centered around 10 mb and 800 mb levels—where the nonstatistical solutions are more accurate than the statistical solutions. The adverse role of statistics is caused by the occurrence of statistically anomalous temperature profile features.

The role of statistics in the temperature profile solution is more dramatically demonstrated on a profile-by-profile basis. Figures 4 and 5 show the individual profiles in which the largest positive differences and the largest negative differences, respectively, of figure 3 occurred. The actual temperature profile (denoted RAOB for radiosonde observation) shown in figure 4 is characterized by a very sharp and cold tropopause. The tropopause layer is less than 100 mb thick and apparently below the resolution of the radiance observations since the nonstatistical solution failed to exhibit this tropopause feature. However, the statistical solution partially reveals the tropopause because this feature commonly exists in temperature profiles at this latitude. The accuracy of the two solutions outside the tropopause region is nearly the same. Both solutions capture the low-level surface inversion and the abrupt change in lapse rate near the 10 mb level.

Figure 5 shows a temperature profile that is statistically anomalous for its latitude. In this situation the statistical solution exhibits erroneous tropopause structure as well as an erroneous stratopause near the 2 mb level. This erroneous structure caused relatively large errors throughout the entire stratosphere. The nonstatistical solution, on the other hand, is in excellent agreement with the actual profile throughout most of the atmosphere.

CONCLUSION

In comparing the statistical and nonstatistical inversion methods, the major conclusion to be drawn is that statistics generally improve the retrieval of the

temperature profile in the tropopause region. In other regions of the atmosphere with the exception of the polar stratosphere, statistics do not appear to significantly improve the accuracy of the temperature profile solution. In statistically anomalous situations, the use of statistics can degrade the solution for the atmospheric temperature profile.

The above conclusions are not necessarily valid when one uses better initial profiles or a different number of spectral radiance observations in the solution. In setting up an operational program to do temperature retrievals, the authors found no significant improvement with statistics when using a 12-hr forecast as the initial profile. On the other hand, we have also found that statistics tend to play a more significant role when only a few radiance observations are available for the temperature profile solution.

DISCUSSION

Chahine: The general iteration method is simply a nonlinear relaxation method written in vector notation. Concerning the minimum information method, or Twomey method, all methods that tend to linearize the Planck function by writing it as a function of frequency multiplied by a function of temperature are accurate *only* if measurements are made in a very narrow frequency range so that we can take the dependence of the Planck function on frequency as the value at the mean point between the two extremes. However we are looking into a complicated atmosphere with clouds, etc., and we would like to get all possible information not only from one frequency range. In that case linearization of the Planck function would not apply.

Falcone: In the microwave region the problem is linear.

Parry: As a meteorologist I have had experience in using these data. Some have been saying that the most probable solution is the best solution. If this were the case for operational purposes, then it would not be necessary to make observations. We could simply use the climatology to make early forecasts. The unusual situation is the important situation.

Smith: The best solution is one that satisfies your measurements, given all the other information you know. This additional information could be contained in statistical covariance matrices or what is generated in dynamical forecast models.

King: Certainly any information, statistical or not, should be considered and used. However, critical is the matter of where you use it. If you feed the statistical information in at the beginning of your algorithm, the results are always a little unconvincing because you don't know whether you are regurgitating what you put in or getting new information. There is also the problem that with the given set of radiances there are infinitely manifold possible solutions. The problem is choosing among one of many, and the statistical criterion has been used as the basis for a uniqueness criterion. This is not the only one. We want some external condition. Another external constraint would be to choose from all of these possible profiles the smoothest one, in the geodesic sense. This is an equivalent radiometric profile, not the actual profile, but it is a completely unbiased one. You may then insert your statistical knowledge to correct that profile.

RECENT DEVELOPMENTS IN THE INVERSION BY THE METHOD OF RELAXATION

Moustafa T. Chahine

Jet Propulsion Laboratory
California Institute of Technology
Pasadena, California 91103

1 N73-11590

ABSTRACT

The relaxation method for inverse solution of the full radiative transfer equation is generalized to solve for all the atmospheric parameters that appear in the integrand as functions or functionals, without any a priori information about the expected solution. Illustrations are presented using the 7.5μ CH_4 band for determining temperature profiles in the Jovian atmosphere, and the 6.3μ band for determining the water vapor mixing ratio in the earth's atmosphere.

INTRODUCTION

The relaxation method of solution [Chahine, 1968, 1970] developed for the determination of a class of atmospheric parameters appearing mainly as *functions* in the integrand of the radiative transfer equation, such as temperature and constant composition mixing ratios, has been generalized [Chahine, 1971] for the determination of the class of parameters that appear as *functionals* in the integrand—for example, variable composition structures.

Here we outline the general nature of the relaxation method, and then conduct several numerical experiments for determining the temperature profile in the atmosphere of Jupiter, and for determining the water vapor mixing ratio profile in the earth's atmosphere. The results discussed here are based in part on Chahine [1971].

THEORETICAL BACKGROUND

The outgoing radiance at a frequency ν , received at level \bar{p} of a plane, parallel atmosphere in local thermodynamic equilibrium, is given by the radiative transfer equation in its integral form as

$$I(\nu) = B[\nu, T(p_s)] \tau(\nu, p_s) + \int_{\ln p_s}^{\ln \bar{p}} B[\nu, T(p)] \frac{\partial \tau}{\partial \ln p} d \ln p \quad (1)$$

assuming a Planck black body boundary condition at the surface p_s with

$$B(\nu, T) = \frac{a\nu^3}{e^{b\nu/T} - 1}$$

The surface emissivity is assumed to be unity, and $\tau(\nu, (q(p)), \dots)$ is the transmittance of a column of absorbers between levels p and \bar{p} .

From an appropriate set of J observed outgoing radiance values $\tilde{I}(\nu_i, \bar{p})$, we aim to determine the unknowns, such as temperature $T(p)$ or composition $q(p)$, by inverse solution of Eq. (1). The extraction of these unknowns from under the sign of integration is difficult because it requires the solution of a nonlinear integral equation with fixed limits. Part of the difficulty can be related to the fact that the mathematical properties of the integrand of Eq. (1) precludes the reduction of this problem to the inverse solution of a linear system of equations

$$\tilde{I}_i = A_{ij}\theta_j \quad (2)$$

A comparison between the properties of Eqs. (1) and (2) reveals that Eq. (2) always has a solution for any arbitrary set of given values of $\tilde{I}(\nu_i)$, while integral equations with fixed limits as Eq. (1) do not always have a solution if $\tilde{I}(\nu)$ is arbitrary. And in this problem \tilde{I} is arbitrary, because it is obtained from measurements that include noise, thereby making the reduction of Eq. (1) to Eq. (2) mathematically improper.

Furthermore, the linear system resulting from Eq. (2) is always highly ill-conditioned; for example, if the unknown on the right-hand side of Eq. (2) is temperature, the condition number of matrix A is 10^5 and 10^7 in the 15μ and 4.3μ CO_2 bands, respectively. Therefore, no temperature profile can be determined from the solution of such a linear system without the use of strong damping functions; when the damping functions are empirical, the results become arbitrary and devoid of any meaning associated with solutions of the radiative transfer equation.

Thus, under the conditions imposed by this problem, a nonlinear approach is mandatory in order to achieve physically and mathematically meaningful results. The relaxation method of solution applied in this paper is nonlinear and is capable of discriminating between noise and true information, without any data smoothing.

The problem to be solved by the method of relaxation reduces to finding values of $T(p)$ or $q(p)$ that satisfy the equation

$$\tilde{I}(\nu_i) = B(\nu_i, T_s)\tau(\nu_i, p_s) + \int_{\ln p_s}^{\ln \bar{p}} B(\nu_i, T(p)) \frac{\partial \tau(\nu_i, p)}{\partial \ln p} d \ln p \quad (3)$$

for all values of $\tilde{I}(\nu_i)$. The solution will be obtained by iteration starting from an arbitrary initial guess ($n = 0$). If the initial guess $T^{(n)}(p)$ or $q^{(n)}(p)$ is the solution, the residuals $R^{(n)}(\nu_i)$ will be satisfied for all frequencies with

$$R^{(n)}(\nu_i) = \frac{\tilde{I}(\nu_i) - I^{(n)}(\nu_i)}{\tilde{I}(\nu_i)} \rightarrow \epsilon \quad (4)$$

But if Eq. (4) is not satisfied we generate a new guess using an appropriate relaxation equation.

To derive the needed relaxation equation we will consider first the simple case when τ is weakly dependent on changes in temperature, which permits the integrand to be treated as a function of $T(p)$. If $T^{(0)}(p)$ does not lead to zero residuals for all values of ν_i , we proceed to satisfy Eq. (3) for each frequency separately. We multiply the right-hand side of Eq. (3) by the ratio $\tilde{I}(\nu_i)/I^{(n)}(\nu_i)$ to get

$$\tilde{I}(\nu_i) - \left\{ \frac{\tilde{I}_i}{I_i^{(n)}} B[\nu_i, T_s^{(n)}] \right\} \tau(\nu_i, p_s) - \int_{\ln p_s}^{\ln \bar{p}} \left\{ \frac{\tilde{I}_i}{I_i^{(n)}} B[\nu_i, T^{(n)}(p)] \right\} \frac{\partial \tau}{\partial \ln p} d \ln p = 0 \quad (5)$$

and generate a new guess for the Planck function from the expression between brackets as

$$B[\nu_i, T^{(n+1)}(p)] = \frac{\tilde{I}(\nu_i)}{I_i^{(n)}} B[\nu_i, T^{(n)}(p)] \quad (6)$$

Next, we pair each frequency ν_i with a specific pressure level p_i , such that a frequency ν_i is designated to sound the atmosphere at the level p_i from which a substantial portion of its upwelling thermal energy originates, and use Eq. (6) as a relaxation equation to determine $T^{(n+1)}(p_i)$ at the J pressure levels. Then, we use a suitable interpolation formula, substitute the new temperature profile into Eq. (3), and repeat the relaxation process until the residuals approach zero.

Equation (6) shows that pairing of frequencies and pressure levels is valid everywhere except for $B = 0$. Therefore, when the given function is exact and continuous the resulting solution should be unique, *if it exists*. However, when $\tilde{I}(\nu)$ is given at a discrete number of frequencies we can no longer obtain a unique solution. Instead, we seek a family of solutions and optimize the results by selecting frequencies that accentuate the physical factors in which we are interested. The selection of a set of sounding frequencies here follows the principles proposed by Kaplan [1959].

GENERAL METHOD OF SOLUTION

Examination of Eq. (6) reveals that the solution of Eq. (3) is obtained by performing, for each frequency ν_i , a different scale transformation on $B^{(n)}(\nu_i, p)$ over the whole range of integration, and then assigning the new value to the preselected pressure level p_i . This suggests an alternative method of solution in which the transformation is applied directly to the unknown variable, in a manner similar to that used by Chahine [1970] for determining the constant mixing ratio. The relaxation equation for temperature then takes the form of

$$T^{(n+1)}(p_i) = \alpha_i T^{(n)}(p_i) \quad (7)$$

where each α_i is a constant scaling $T^{(n)}(p)$, to satisfy Eq. (3) for each frequency separately. This general approach is required when the integrand is a functional of the unknown parameters as in the case of determining composition profiles, and the case of determining temperature profiles when the transmittance is strongly dependent on temperature variations. Obviously, when τ is weakly dependent on T , the scaling constants α_i for temperature can be obtained directly from Eq. (6).

To apply the general approach for determining composition profiles, let us first integrate Eq. (1) by parts and, for convenience, write the results as

$$I(\nu, \bar{p}) = B(\nu, \bar{p}) - \int_{\ln p_s}^{\ln \bar{p}} \tau(\nu, \langle q(p) \rangle, \dots) \frac{\partial B}{\partial \ln p} d \ln p \quad (8)$$

where $q(p)$ is the composition profile, and the notation $\langle q(p) \rangle$ indicates that the transmittance is a functional of $q(p)$, since τ depends on the distribution of $q(p)$ between p and \bar{p} .

To determine $q(p_i)$ from a given set of radiance measurements, assuming that the temperature profile is known, we pair each frequency ν_i with a pressure level p_i , then make an initial guess for $q^{(n)}(p_i)$ and solve the radiative transfer equation

$$\tilde{I}(\nu_i) - \bar{B}(\nu_i) + \int_{\ln p_s}^{\ln \bar{p}} \tau(\nu, \langle \alpha_i q^{(n)}(p) \rangle, \dots) \frac{\partial B}{\partial \ln p} d \ln p = 0 \quad (9)$$

to obtain the scaling constants α_i corresponding to each sounding frequency ν_i (see appendix). We generate the following iterations through the relaxation equation

$$q^{(n+1)}(p_i) = \alpha_i q^{(n)}(p_i) \quad (10)$$

The iteration process is repeated until each value of the scaling constants approach unity, which is equivalent to satisfying the residuals in Eq. (4). This relaxation method of solution leads to accurate determination of composition profiles without any a priori information about the expected solution, and the residuals converge to zero if, and only if, $q^{(n)}(p)$ is the solution.

The process of optimizing the results obtained from a discrete set of sounding frequencies should not violate the law of conservation of information, so that from a set of J sounding frequencies a maximum of J scaling constants α_i can be determined. The use of weighted scaling constants of the form

$$\bar{\alpha}_i = \frac{\sum_{j=1}^J \alpha_j W(\nu_j, p_i)}{\sum_{j=1}^J W(\nu_j, p_i)} \quad (11)$$

in Eqs. (7) and (10) can be useful in the presence of large errors in measurement. In Eq. (11) $W(\nu_j, p_i)$ is the fractional contribution from the i th pressure level to the total outgoing radiance at frequency ν_j . An approximation to the values of the weights has been given by Smith [1970] for temperature

$$W(\nu_j, p_i) = \frac{\partial \tau(\nu_j, p_i)}{\partial \ln p} \quad \text{for } p_i < p_s$$

and

$$W(\nu_j, p_s) = \tau(\nu_j, p_s) \quad (12)$$

and for composition profiles as

$$w^{(n)}(\nu_j, p_i) = \left[\frac{\partial \tau(\nu_j, p_i)}{\partial \alpha} \right]^{(n-1)} \frac{\partial B(\nu_j, p_i)}{\partial \ln p} \quad (13)$$

NUMERICAL EXPERIMENT

The general relaxation method has been applied to invert synthetic radiance data generated on a computer from two frequency ranges. The numerical experiment is intended to demonstrate the accuracy and stability of solutions derived without any a priori information about the expected profiles.

The 7.5 μ CH₄ Band of Jupiter

The set of seven sounding frequencies shown in figure 1 has been selected by Taylor [1971] to sound the temperature profile above the pressure level of 1 atm ($Z = 36$ km) in the Jovian atmosphere. Assuming a constant mixing ratio of methane, and the model temperature profile shown in figure 2, a set of seven simulated values of $\tilde{T}(\nu_i)$ were generated and used to reconstruct the model temperature profile starting from three different initial guesses:

1. $T^{(0)}(p) = 100.0$
2. $T^{(0)}(p) = 200.0 - 0.01 (Z - 80)^2$
3. $T^{(0)}(p) = 100.0 + 0.01 (Z - 80)^2$

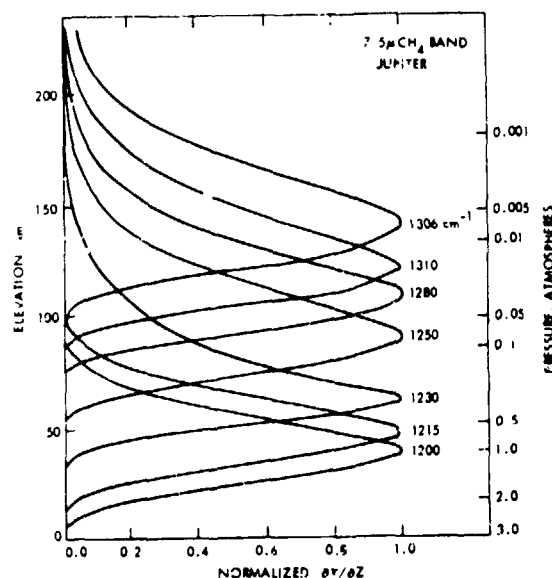


Figure 1.— Normalized weighting functions $d\tau/d \ln p$ for the selected set of 7 sounding frequencies in the 7.5 μ CH₄ band of Jupiter. The instrumental slit function is 5 cm⁻¹ for frequency 1306 and 20 cm⁻¹ for the rest. [After Taylor, 1971].

The transmission function was recomputed during each of the first four iterations, to correct for the large changes in temperature from $T^{(0)}(p)$ to $T^{(4)}(p)$. Therefore, during the first four iterations the integrand was treated as a functional of $T(p)$ and the general relaxation method was applied; the corresponding scaling factors α_i were evaluated numerically by linear interpolation. The solution converged rapidly and uniformly so that after the fourth iteration the values of α_i were derived directly, without the need for interpolation, from the Planck function ratio of Eq. (6), as

$$\alpha_i = \frac{b\nu_i / T^{(n)}(p_i)}{\ln \left(1 - \left[1 - \exp \left[b\nu_i / T^{(n)}(p_i) \right] \right] \left[I_i^{(n)} / \tilde{I}_i \right] \right)}$$

Figure 3 shows the dependence of the final answer on the initial guess. The span of the thin horizontal bar can envelope the whole family of solutions obtained from these three widely varying initial guesses.

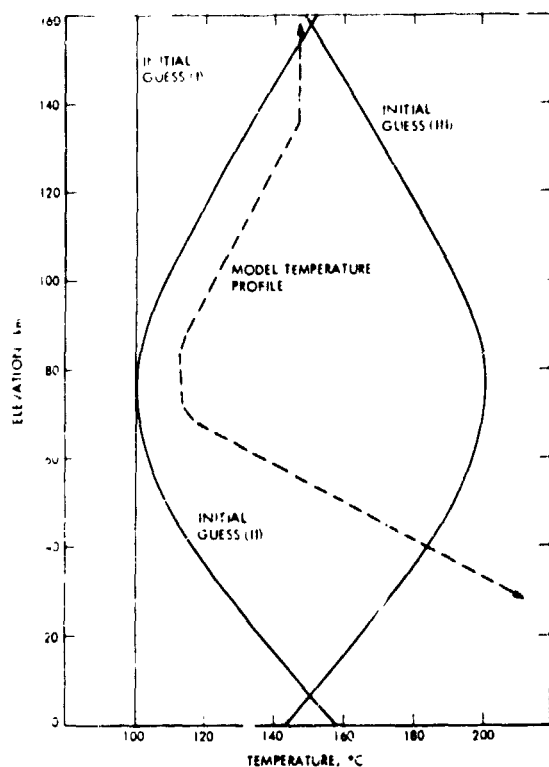


Figure 2.— Three different initial guesses used to start the iterative solution for the determination of the model temperature profile of Jupiter.

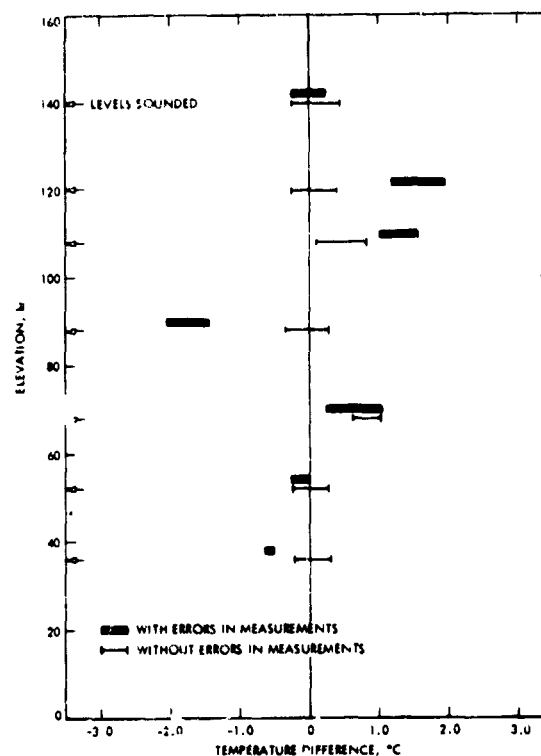


Figure 3.— Effect of the initial guess on the final solution for the cases of no errors and with errors in measurements.

To study the effect of measurement errors on the final solution a set of simulated large errors were introduced to the set of exact radiance values. The errors were given a random distribution of signs and a maximum value of 9.6% with an rms value of 5.42%. The corresponding temperature profile was reconstructed as in the case of zero errors. The results, shown as heavy bars in figure 3, clearly demonstrate the weak dependence of the final answer on the initial guess.

The use of weighted scaling factors

$$\bar{\alpha}_i = \frac{\sum_{j=1}^J \alpha_j \partial \tau(v_j, p_i) / \partial \ln p}{\sum_{j=1}^J \partial \tau(v_j, p_i) / \partial \ln p}$$

did not alter the solution. However, in the case of no errors in measurements the rate of convergence became much slower; but in the case of large errors in measurement the use of $\bar{\alpha}_i$ improved the results and removed any ambiguity about the terminal order of iteration. Figure 4 shows that the solution for the case of 5.42% errors changed by less than 0.1° from iteration 11 to 25. The iteration process should be terminated when the residuals reach their

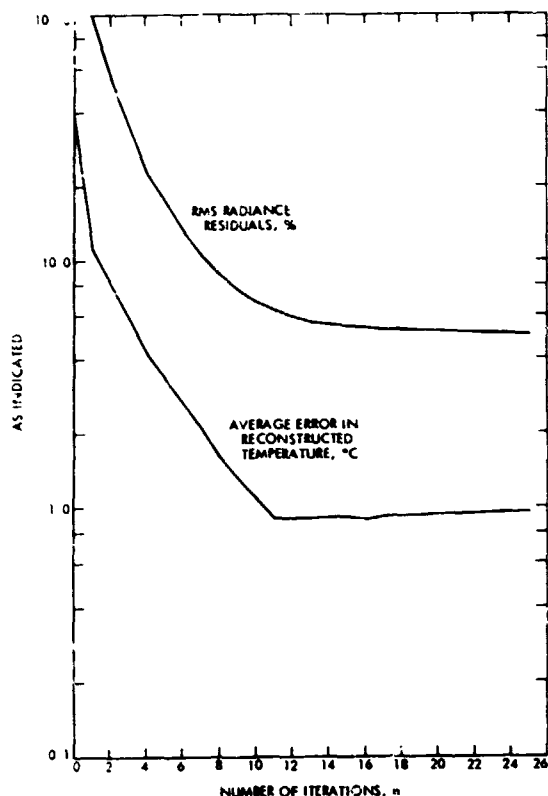


Figure 4.— Variations of the rms residuals with iterations in comparison with the rate of convergence of the temperature profile, using weighted scaling factors the process or relaxation.

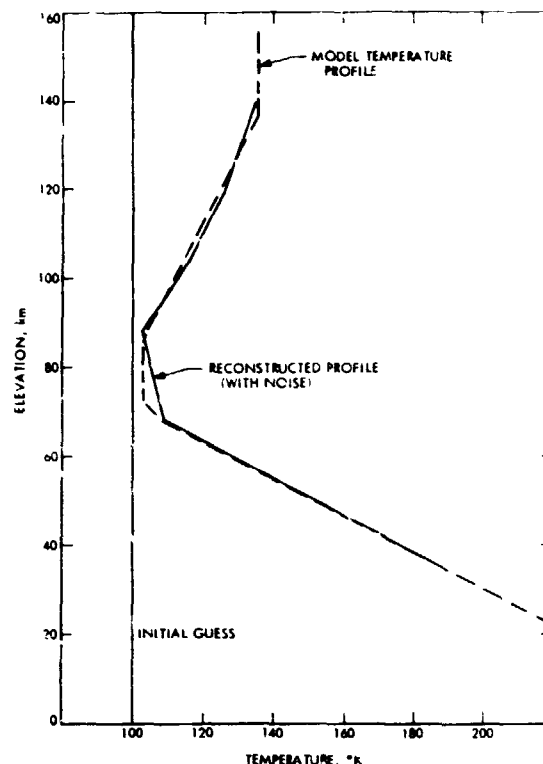


Figure 5.— Comparison between exact and reconstructed profiles starting from an initial guess $T^{(0)}(p) = 100.0$ and with 5.24% rms errors in radiance measurements.

asymptotic value, provided that quadrature and truncation errors are much smaller than errors in τ and I_i . Figure 5 shows a typical comparison between the exact temperature profile and the reconstructed temperature values with linear interpolation, for the case of 5.42% random errors in \tilde{T} with initial guess (1).

6.3μ Water Vapor Band in the Earth Atmosphere

A set of sounding frequencies shown in figure 6 was selected for determining the water vapor mixing ratio profile from simulated observations in the earth's atmosphere. A set of synthetic radiance data corresponding to the mixing ratio profile shown in figure 7 was generated for the case of the U.S. Standard Atmosphere (1962) temperature profile. The computation of $I(v_i)$ from Eq. (1) was carried out using a modified Simpson's rule with a first-order interpolation formula on the scale of figure 7; the relative error caused by numerical quadrature is less than 0.47%.

These simulated radiance values were used to reconstruct the composition profile starting with an initial guess $q^{(0)}(p) = 0.1 \text{ g/kg}$. Because the transmittance is a functional of $q(p)$, the values of the scaling constants α_i were determined by interpolation throughout the iteration process.

The results illustrated in figure 7 show that the relaxation method is capable of reconstructing the composition profiles from radiance observations alone without any a priori information about the solution. The selection of different sets of sounding frequencies from the same spectral range, and the use of water vapor mixing ratio $q^{(0)}(p) = p, p^2$, or p^3 as initial guesses did not change the accuracy of the results. The solid bar is large enough to envelop the whole family of reconstructed solutions.

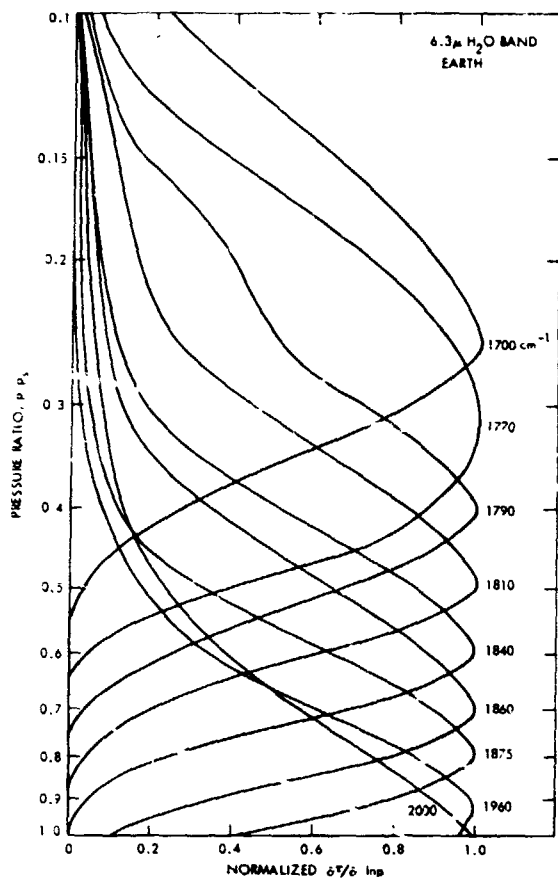


Figure 6.— Normalized weighting functions $dr/d \ln p$ for the selected set of 9 sounding frequencies computed for an instrumental slit function of 5 cm^{-1} .

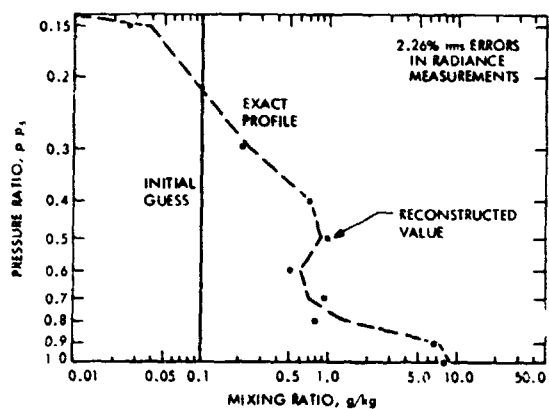


Figure 8.— Reconstructed mixing ratio profile from radiance measurements with 2.24% rms random errors.

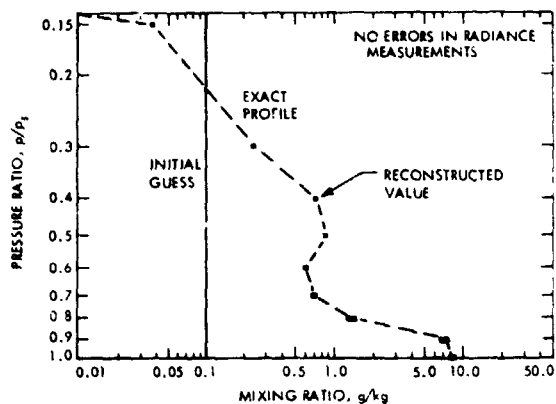


Figure 7.— Reconstructed mixing ratio profile from exact synthetic radiance measurements.

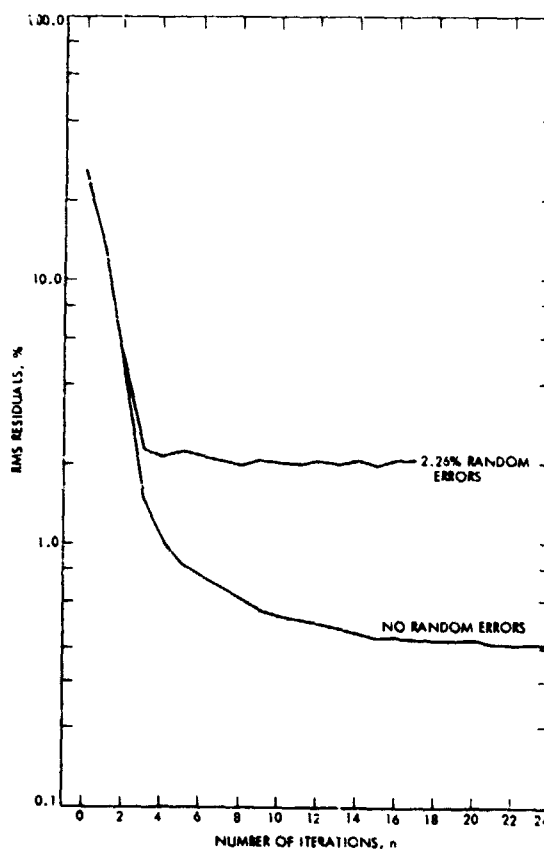


Figure 9.— Rate of convergence of solutions as judged by the variations of residuals with number of iterations, using weighted scaling factors in the process of relaxation. The order of iteration should be terminated when the residuals approach their asymptote, which corresponds to a cutoff at or after iteration 24 and 10 for zero and 2.26% errors, respectively.

The effect on the reconstructed profiles of random errors in radiance measurements is shown in figure 8. This effect is more pronounced in the lower atmosphere; however, it is a characteristic of the spectral range used in this illustration and should not necessarily be true in other frequency ranges.

The rate of convergence is determined from the behavior of the residuals. Figure 9 shows that $\langle R^{(n)} \rangle_{\text{rms}}$ decreases rapidly and then approaches an asymptotic value approximately equal to the sum of the averaged random and numerical errors. The iteration process should be terminated when the rms residuals approach this asymptote.

Application of the weighted scaling constant of Eq. (11) reduced the efficiency of the relaxation process for the case of zero errors, but improved the stability of the solution in the presence of large random errors in measurement, and removed any ambiguity about the number of needed iterations. In the case of the upper curve in figure 6 the optimum profile can be obtained anywhere from iteration 10 to iteration 35 with the same degree of accuracy.

CONCLUSION

The efficiency of any fixed set of sounding frequencies can be strongly dependent on atmospheric conditions, particularly in the presence of a variable composition mixing ratio. An improvement in the level of efficiency can be obtained by the use of weighted scaling factors, in the presence of large errors in measurement.

The relaxation method can be applied in conjunction with any interpolation formula. The extent of interpolation is dictated by the quadrature requirements, and by the need to optimize the quality of solutions obtained from a given set of sounding frequencies.

Finally, the method of relaxation, which was applied in this paper to study homogeneous nonscattering atmospheres, can be applied also to the study nonhomogeneous scattering atmospheres without any loss of generality.

ACKNOWLEDGMENT

This paper presents the results of one phase of research carried out at the Jet Propulsion Laboratory, California Institute of Technology, under Contract Number NAS 7-100, sponsored by the National Aeronautics and Space Administration.

APPENDIX

INTERPOLATION PROCESS FOR DETERMINING α_i

The determination of α_i from Eq. (9) can be accomplished with almost any standard method for solving non-linear equations.

Let

$$I(\nu_i, \alpha_i) = B(\nu_i, \bar{p}) - \int_{\ln p_s}^{\ln \bar{p}} \tau(\nu_i, \langle \alpha_i q(p) \rangle, \dots) \frac{\partial B}{\partial \ln p} d \ln p$$

and let

$$r(\nu_i, \alpha_i) = \tilde{I}(\nu_i) - I(\nu_i, \alpha_i) \rightarrow \epsilon$$

where ϵ is the magnitude of numerical quadrature errors. We aim to determine α_i by interpolation beginning with

$$\alpha_i^{(1)} = 1.0$$

We compute $r_i^{(1)} = r(\nu_i, \alpha_i^{(1)})$ and use as a second interpolation point, say, the value of

$$\alpha_i^{(2)} = \frac{\tilde{I}(\nu_i)}{I[\nu_i, \alpha_i^{(1)}]}$$

we compute $r_i^{(2)}$ and obtain $\alpha_i^{(k)}$, $k = 3$, by linear interpolation as

$$\alpha_i^{(k)} = \alpha_i^{(k-2)} - r_i^{(k-2)} \frac{\alpha_i^{(k-2)} - \alpha_i^{(k-1)}}{r_i^{(k-2)} - r_i^{(k-1)}} \geq 0$$

The next value $\alpha_i^{(k+1)}$ is obtained by linear interpolation between $r_i^{(k)}$ and the smallest of all other $|r_i|$ values. The root α_i corresponds to the minimum of all $|r_i^{(k)}|$.

A check for the possible existence of multiple roots is advisable, particularly in the presence of strong inversion in the temperature profile. However, if the wrong root is selected, the residuals of Eq. (4) will diverge.

LIMB RADIANCE INVERSION:
ITERATIVE CONVERGENCE FOR A NONLINEAR KERNEL

John C. Gille[†]
Department of Meteorology
and
Geophysical Fluid Dynamics Institute
Florida State University, Tallahassee

N73-11591

ABSTRACT

The features of the limb inversion problem that differentiate it from the nadir problem are described; the most important of these are the sharply peaked weighting functions and the nonlinear kernel that arises because the atmosphere is required to be in hydrostatic balance. The problem may be split into two parts—given the pressure at one level and the measured radiances, can the temperatures be retrieved? And how can the pressure be obtained at one level? The second problem is shown to depend on the first, which is solved iteratively. The convergence properties and final accuracies are presented. Inversion to obtain trace constituents is briefly discussed.

INTRODUCTION

Previous writers have described methods for inverting measurements of the upwelling radiance in several narrow spectral intervals. The radiometer in those cases is pointed close to the nadir. Limb radiance measurements are made while a radiometer is scanned from the planet to space across the planetary limb. To the extent that thermal radiation emitted by the planetary atmosphere is passively received by a radiometer, there are similarities between the limb- and nadir-viewing problems. However, the change in geometry changes the inversion problem, most importantly by making the weighting functions or kernels of the integral equation strongly dependent on the temperature profile. Here we describe the results of applying a version of Chahine's iterative algorithm to this nonlinear problem and note some of the technique's convergence properties.

Techniques for inverting limb measurements are not as developed as those for nadir observations, and experience much more limited. Although the first satellite inversion scheme [King, 1956] proposed an angular scan as a means of probing the entire atmosphere, Kaplan's [1959] suggestion for a frequency scan at a given viewing angle has been exploited for practical reasons. Schemes for inverting measurements of limb radiance have been put forward since the mid-1960s, but only in the last year has an algorithm been developed to use data that are easily measured with state-of-the-art technology. The work reported here should therefore be regarded as an initial investigation.

THE INVERSION PROBLEM

The geometry and associated nomenclature of limb viewing are shown in figure 1. The radiometer receives radiation emitted along a particular path, which may be identified by the tangent height (h), the altitude of the lowest point (tangent point) above the surface. The atmosphere is sounded by receiving radiation from different tangent heights, from $h < 0$ (ray paths intersecting the surface) to $h \sim 100$ km, from which a very small signal is received.

[†] Present affiliation: National Center for Atmospheric Research, Boulder, Colorado.

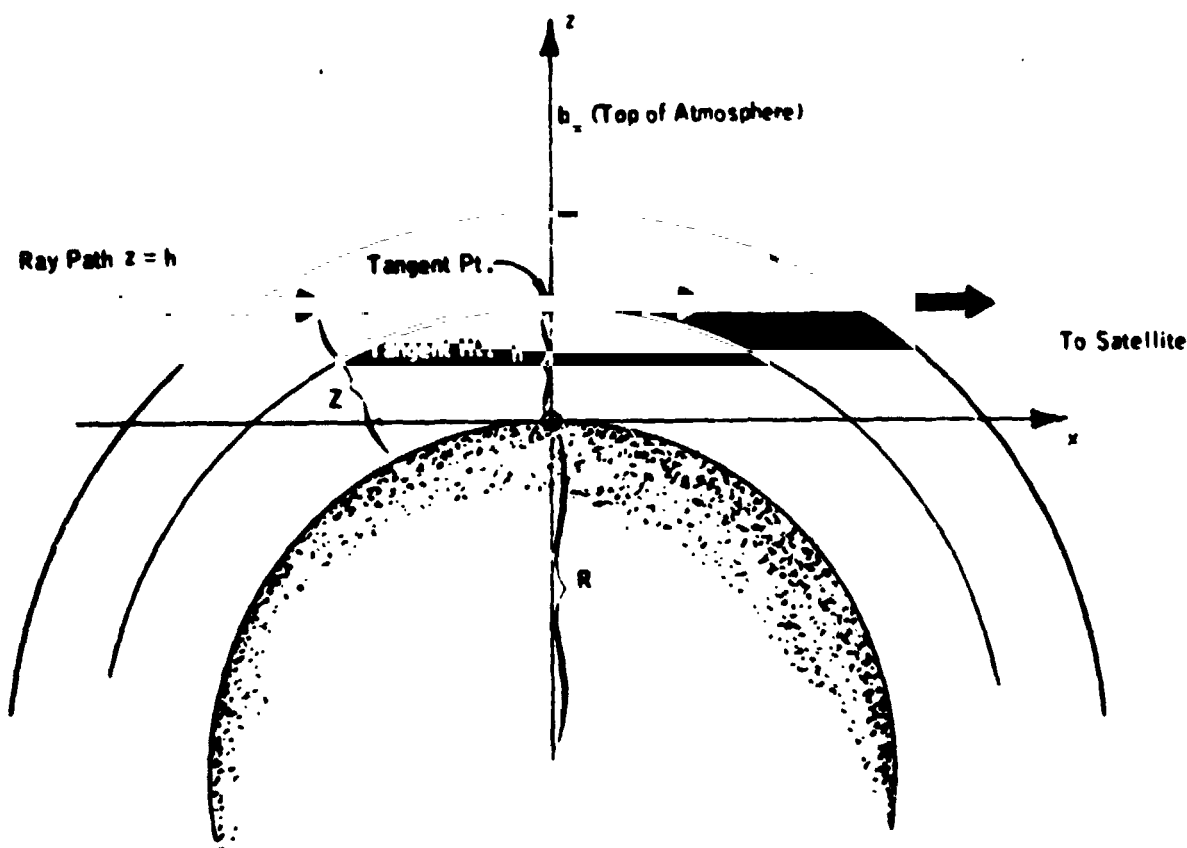


Figure 1.- The geometry of limb viewing, indicating nomenclature. A profile is obtained by measuring outgoing radiance as a function of tangent height.

Some features of the inversion problem are immediately apparent from figure 1. A relatively minor point is that all the signal originates in the atmosphere; the boundary term in the equation of radiative transfer vanishes, since the far side of the atmosphere is cold space. Although this is not too important for temperature determination, it is an enormous aid in the determination of minor atmospheric constituents, such as ozone and water vapor.

We also note that there is much more emitting material along a horizontal path than on a vertical path through the atmosphere. This indicates that we may get a measurable signal from high levels, but that the large opacity at low levels will not allow signals from the tangent point to reach the radiometer. The problem is compounded by clouds. For these reasons, limb scanning is not a good way to probe below the upper troposphere. Although lower regions can be sounded, subsequent results are presented only above 15 km.

More important is the inherent high vertical resolution. From the geometry, it is clear that none of the signal originates below the tangent point. If the atmosphere above the tangent point is divided into a series of thin shells of equal thickness, the longest horizontal segment of the ray will lie in the shell immediately above the tangent point. Because density falls off exponentially with height, this shell will also contain a large fraction of the mass along the path. For a wide range of tangent heights, the largest part of the emitted radiation will originate in this shell. Thus, if we write an expression for the outgoing radiation I when viewing tangent height h as

$$I_i(h) = \int_0^{\infty} B_i[T(z')] W_i(h; z') dz' \quad (1)$$

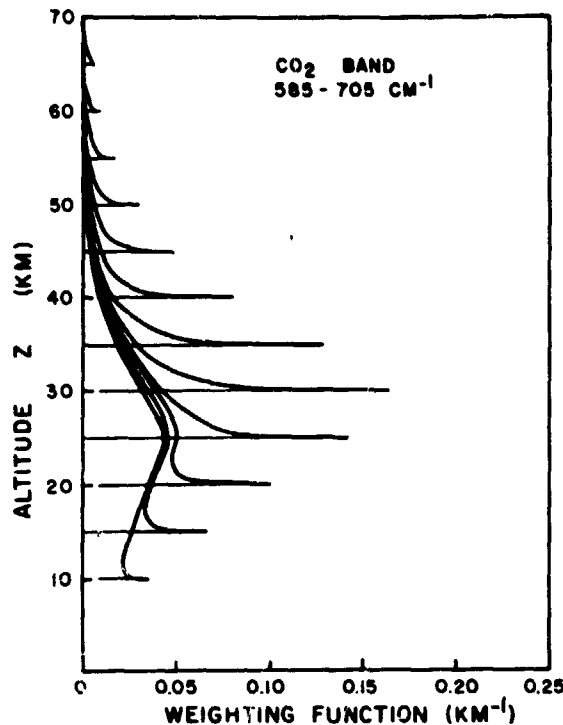


Figure 2.— Limb viewing weighting functions $W(h; z')$ (defined in Eq. (1)) for the ideal case of an instrument with an infinitesimal vertical field of view. The strongly peaked shape is an important difference from the broader nadir-viewing weighting functions.

in which subscript i denotes the spectral interval, B the Planck intensity, T the temperature, and z' the altitude, then $W(h; z')$ is a weighting function, indicating how much level z' contributes to the intensity received when viewing h .

Weighting functions for several tangent heights through a midlatitude atmosphere are shown in figure 2 for a broad band covering most of the $15 \mu\text{m}$ band of CO_2 with an infinitesimal vertical field of view. For tangent heights above 25 km a large part of the contribution comes from within about 3 km of the tangent height. These should be compared to the much broader nadir weighting functions discussed in the preceding sections of this chapter. Below 25 km, the weighting functions begin to look like the broad nadir-viewing weighting functions, although there is still a spike at the tangent point due to the less strongly absorbing wings of the band.

The narrow weighting functions give an indication of the potential vertical resolution. A real instrument will have a finite field of view, and its weighting functions, convolutions of ones like those in figure 2, will be broader. (This immediately indicates that if one is to use the possible inherent vertical resolution the vertical instrument field of view at the horizon must be less than about 6 to 8 km to improve on nadir viewing.) However, spectral resolution is not necessary, and a broad spectral interval may be used.

The most important difference comes about because of the sensitivity of the weighting functions to the amount of emitting material, especially at the tangent point, and the transmission along the path. For temperature determinations we shall measure the radiation emitted by a uniformly mixed constituent, carbon dioxide. This means that the weighting functions will be sensitive to the atmospheric density profile, which in turn depends on the temperature profile through the hydrostatic equation. Simply stated, higher temperatures in the lower part of the atmosphere move density surfaces (or, approximately, surfaces of constant optical depth) upward. Because of the sharp weighting functions, even a shift of a given density surface by 1 km can appreciably modify the weighting functions. Thus, instead of Eq. (1) we write

$$I_i(h) = \int_h^\infty B_i[T(z')] W_i[T(z); h; z] dz' \quad (2)$$

To develop the specific expression to be inverted, we note from the geometry of figure 1 that the outgoing radiance from a nonscattering atmosphere in local thermodynamic equilibrium may be written

$$I_i(h) = \int_0^\infty B_i[T] \frac{d\tau_i(h; x)}{dx} dx \quad (3)$$

where τ is the transmission between x and the spacecraft (∞) along the path through h , x is the distance coordinate along the ray path with origin at the tangent point, and the other symbols are as above.

For a finite band in the spectrum

$$\tau(h; x) = \tau(h; \bar{a}, \bar{p}) \quad (4)$$

in which \bar{a} , \bar{p} are the temperature weighted absorber amount and temperature-absorber weighted pressure, respectively [Goody, 1964; Rodgers and Walshaw, 1966].

Incorporating Eq. (4) into Eq. (3), converting to a vertical integral, using the hydrostatic law and perfect gas law, and specializing to a uniformly mixed constituent, we obtain [Gille and House, 1971].

$$I(h - z_0) = p_0 \int_{h-z_0}^{\infty} B[T(z')] \left| \frac{dx}{dz} \right|_{z'} \left[a[T(z')] \left\{ -\left(\frac{\partial \tau}{\partial \bar{a}} \right)_a - \left(\frac{\partial \tau}{\partial \bar{a}} \right)_p \right\} - b[T(z')] \right. \\ \left. \left\{ \left(\frac{\partial \tau}{\partial \bar{p}} \right)_a - \left(\frac{\partial \tau}{\partial \bar{p}} \right)_p \right\} \right] \exp \left[- \int_{z_0}^{z'} CT(z)^{-1} dz \right] T(z')^{-1} dz' \quad (5)$$

where p_0 is the pressure at an arbitrary level z_0 , C is a constant, and $a(z')$, $b(z')$ contain constants and the information on the temperature variation of the transmission. Subscripts a and p denote the z anterior and posterior to the tangent point. The exponential term relates the density at z' to that at z_0 through the hydrostatic equation. Although not written explicitly, the transmission derivatives depend on density in the ray paths, and consequently on the hydrostatic equation as well.

Equation (5) makes use of the fact that all other altitudes may be measured relative to z_0 . This is extremely important, because it is instrumentally far easier to determine relative altitudes than absolute altitudes. Thus, given a measured radiance profile like the broad channel measurement in figure 3, but with no attached altitude scale, we may pick a particular point on the curve such as A , designate its height as the reference height z_0 , and measure all other heights relative to it. Henceforth, for convenience we designate relative tangent heights and altitude by asterisks: $h - z_0 = h^*$, $z - z_0 = z^*$.

Equation (5) defines the inversion problem. We have measurements of $I(h^*)$, and wish to

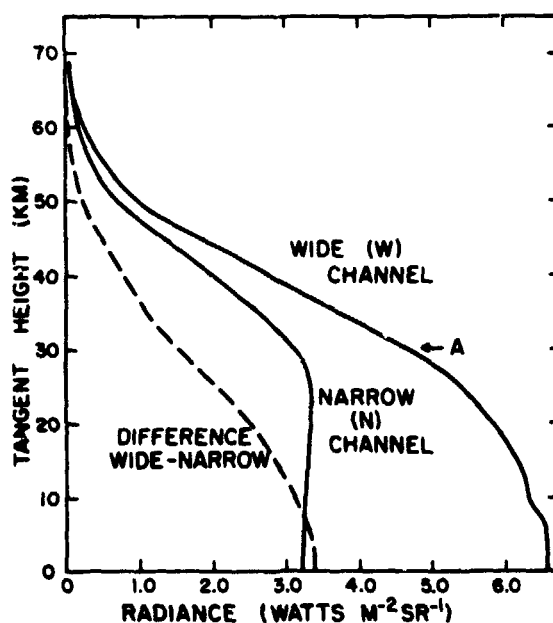


Figure 3.- Limb radiance as a function of tangent height for wide (585-705 cm^{-1}) and narrow (630-685 cm^{-1}) channels in the 15 μm band of carbon dioxide. Their difference results from emission by generally weaker lines away from the band center, and provides information on deeper atmospheric layers. The arrow (A) indicates the arbitrary level whose altitude is designated z_0 where the pressure p_0 is required. (See Description and Results of an Iterative Inversion Algorithm.)

determine $T(z^*)$. Clearly p_0 is a critical parameter. The inversion problem may then be separated into two parts:

1. Given p_0 and measurements of $I(h^*)$, how may $T(z^*)$ be determined?
2. How may p_0 be determined?

The first question is treated below. The second, closely related to the first, is discussed in a later section.

Description and Results of an Iterative Inversion Algorithm

The complex nonlinear manner in which the temperature enters on the right-hand side of Eq. (5) suggests that some type of iterative method will be required. Basically, the purpose is to bring about temperature fields that satisfy both the radiative transfer and the hydrostatic equations. The sharply peaked weighting functions indicate that discrepancies between observation and calculation at a particular tangent height may be used to correct the temperature at that elevation, as in *Chahine's* [1968, 1970] method.

The steps in such a solution are as follows:

1. In the first iteration, an initial temperature profile is assumed.
2. The temperature profile is used with p_0 and the hydrostatic equation to distribute the atmospheric mass in altitude. The importance of this step is the major difference from a nadir-viewing algorithm.
3. The outgoing radiation is calculated as a function of tangent height from one of the equivalent Eqs. (2), (3), and (5).
4. The calculated outgoing radiances are compared to the observed values, tangent height by tangent height. If a convergence criterion is satisfied, the temperature profile is accepted as the solution. A typical convergence criterion is that the rms difference between the calculated and observed radiances over the profile (the radiance residuals), be slightly less than the noise of the radiometric measurements.
5. If the convergence criterion is not met, the discrepancies between the calculated and observed radiances at each tangent height are used to adjust the temperature there.

Making the assumption that the discrepancy is due to errors in Planck radiance alone, a corrector or relaxation equation is [Gille and House, 1971].

$$T^{n+1}(h^*) = T^n(h^*) + \frac{2T^n(h^*)^2}{c_2 \bar{\nu}} \frac{I(h^*) - I_c^n(h^*)}{I(h^*) + I_c^n(h^*)} \quad (6)$$

in which superscripts denote the order of iteration, c_2 is the second radiation constant, $\bar{\nu}$ is a mean frequency, I is the observed radiance, and I_c is the computed radiance value. The improved temperature is then used in step 2, and the procedure iterated until the convergence criterion is met.

Results of applying this procedure to data synthesized for a broad spectral interval covering much of the $15 \mu\text{m}$ band of CO_2 for an actual midlatitude winter sounding are shown in figure 4, in which the solid line is the original sounding. The dotted line is an inversion of ideal data—that is, noise free data having an infinitesimal field of view at the horizon. The inverted points were spaced 1 km apart vertically. The rms radiance residuals were reduced only to $0.01 \text{ Wm}^{-2} \text{ sr}^{-1}$, which was taken as a rough instrumental noise figure. The rms temperature error is 0.91°K , and the solution reproduces the initial profile very well.

Note that the thin warm anomaly at 37 km is retrieved quite accurately. Smaller variations below 25 km are not followed as well.

The dashed line indicates the results obtained when data for the same atmosphere, but now including the effects of a finite field of view and random noise of $0.01 \text{ W m}^{-2} \text{ sr}^{-1}$. Now the rms temperature error is 4.2°K from 15 to 60 km, and 2.43°K from 15 to 54 km. In this case, the thin features are not retrieved very well (there is just a hint of the 37 km anomaly), but the overall curve is followed quite closely. Large unreal oscillations occur at the top of the sounding.

The spacing of the inverted points is not fixed. We may space them 3 km apart instead of 1 km, in which case the results are similar to those for 1 km spacing, except that the large oscillation at the top of the sounding do not appear.

CONVERGENCE PROPERTIES

It is very instructive to follow the convergence to a solution. Figure 5 displays the reduction of the radiance residuals as a function of iteration number for the two cases discussed above. Also shown are radiance residuals when the inverted points are spaced 3 km apart, and intermediate points are obtained by linear interpolation. The 3 km calculations require much less computer time.

This is a very instructive plot. Note first that for noisy or ideal data, 1 or 3 km spacing, the residuals are nearly identical for the first nine iterations. Beyond that point, the residuals of the ideal data continue to drop rapidly, while the noisy data residuals level off. The operational significance is important: the first stages of an inversion may be done rapidly with widely spaced points, and later stages may be done with higher resolution.

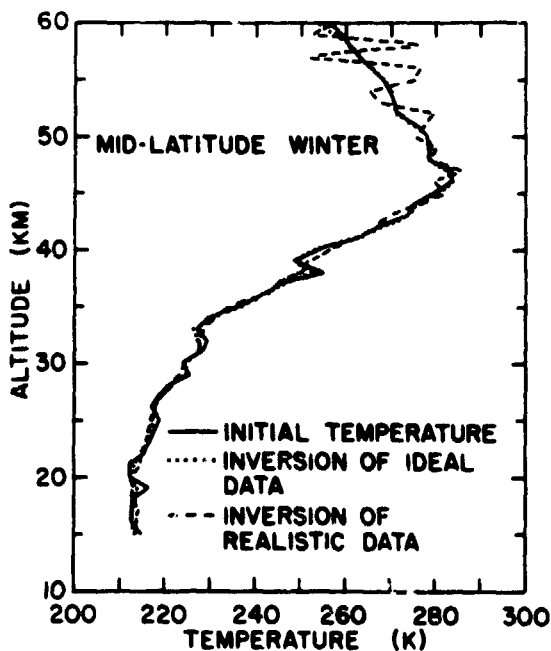


Figure 4.— Solutions obtained by inverting synthesized ideal data (no noise, infinite vertical resolution), and realistic data ($0.01 \text{ W m}^{-2} \text{ sr}^{-1}$ random noise, 2 km field of view), compared to an actual midlatitude winter sounding.

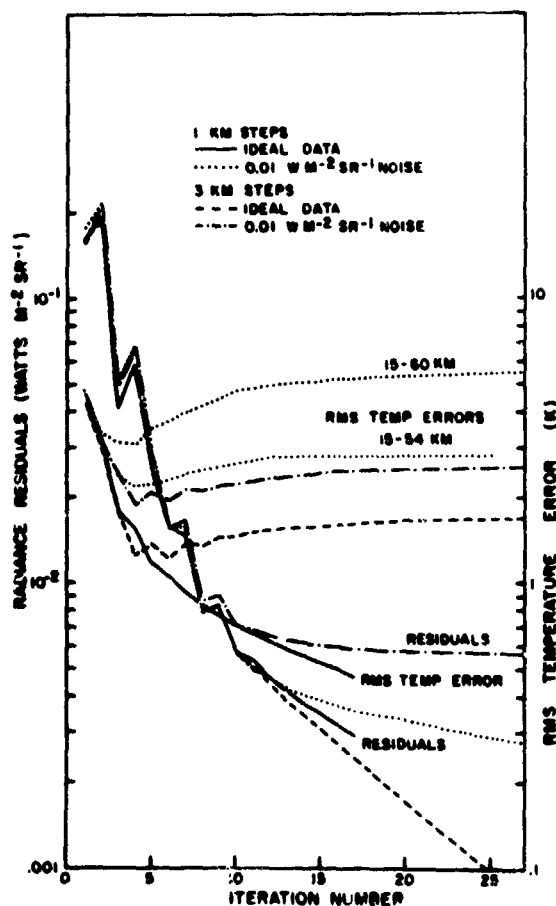


Figure 5.— Root-mean-square radiance residuals and temperature errors as functions of iteration number, for the midlatitude winter profile for ideal and realistic data, 1 and 3 km spacing of the inverted values.

The second notable point is that for the first several iterations there is an oscillatory behavior with small residual increases being followed by large decreases. This reflects the up and down sloshing of the atmosphere, caused by application of the hydrostatic equation at each step.

The behavior of the rms temperature errors also is shown in figure 5. For ideal 1 km data, the temperature error drops monotonically, reaching 0.3°K after 17 iterations. For the ideal 3 km data, it drops for four iterations, reaches a minimum of 1.2°K at the sixth iteration, and then slowly rises to 1.7°K. At the eighth iteration, when the residuals have dropped below the noise ("convergence"), the error is 1.35°K. Clearly, in this case the convergence criterion is letting the solution go past the optimum point. After the fourth iteration, the 1 km inversion is always better than the 3 km inversion.

For data containing $0.01 \text{ W/m}^2 \text{ sr}$ noise, the rms temperature errors for both 1 and 3 km inversions pass through minima before the radiance residuals reach the noise level. In this particular case, the 3 km temperature error reaches a minimum value of 1.9°K at the fourth iteration, and only 1.55°K from 15 to 54 km. At the eighth iteration, when the radiance residuals have dropped below the noise, the temperature errors are 2.1°K for both 15 to 54 and 15 to 60 km. Although further iteration reduces the rms radiance errors from 8.8×10^{-3} to 5.6×10^{-3} , the rms temperature error increases to 2.5°K, or 2.4°K from 15 to 54 km.

For noisy data, the results of 1 km spacing of inverted points is worse. Now the 15 to 60 km error is 3.2°K at the fourth iteration, rising to 4.2°K at the eighth, and increasing to about 5.5°K with further iteration. If we restrict our attention to the rms temperature errors from 15 to 54 km, we again find a minimum of 2.2°K at the fourth iteration,

increasing to 2.5°K at the eighth, and 2.8°K at iteration 25. As could be seen in figure 4, with 1 km spacing a great deal of the error is accumulated at the top of the sounding; the rms error is much improved by ignoring this region. Comparison of the inversion data and the 15 to 54 km data with 1 km spacing indicates that we have lower accuracy in the latter case, although higher resolution. This is analogous to the use of larger or smaller values of γ in Twomey's [1963] method (see also earlier sections in this chapter by Mateer and by Conrath and Revan) or points on a Backus-Gilbert type curve, which depicts the tradeoff of accuracy versus resolution explicitly (see the discussion by Parker in this chapter).

For any iterative scheme, and especially for a strongly nonlinear problem, we expect the rate of convergence and perhaps the final accuracy to depend on the initial guess. The results of inverting a tropical profile starting from the midlatitude standard atmosphere and from the tropical standard atmosphere are shown in figure 6. Note that for the standard atmosphere start, 18 iterations are necessary for noise-free data and 21 for noisy data. The tropical start residuals drop much more rapidly, reaching the noise level in 9 and 11 iterations.

This suggests that operationally one might try several initial profiles to discover the one with the lowest residuals, or the most rapid drop in residuals after one iteration, to get the best starting profile.

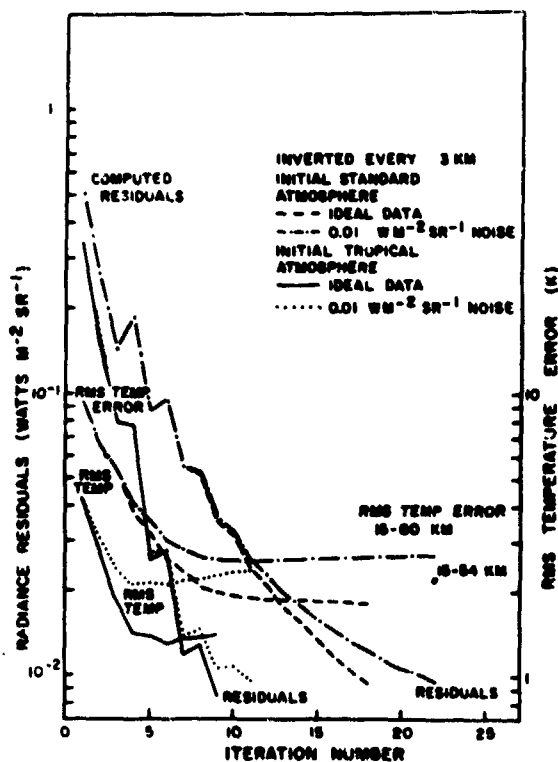


Figure 6.— Root-mean-square radiance residuals and temperature errors as functions of iteration number for a tropical atmosphere. The effects of starting from a midlatitude standard atmosphere or the tropical standard atmosphere are shown with ideal and realistic data, with inverted levels spaced every 3 km.

Except for the noise-free data on the standard atmosphere start, all the rms temperature errors go through a minimum, although the standard atmosphere minimum is very shallow. Error at "convergence" for the noisy data is 2.65°K from 15 to 60 km for the standard atmosphere start and 2.35°K for the tropical start, up from 2.1°K at the fourth iteration.

Consideration of the retrieved temperature profiles (not shown) indicates that with the size of the instrumental noise used, there is difficulty in retrieving the position of the tropical tropopause (approximately 18 km) and the temperature distribution below it. Thus, errors at the top and bottom of the 15 to 60 km region are larger than in the intermediate region, as of course we might expect from the weighting functions.

DETERMINATION OF p_0

In the study so far, we have assumed that p_0 is known. Now we shall show how this problem may be solved, if the preceding one is. Figure 7 shows the effects of inferring a temperature profile (in this case the U.S. Standard Atmosphere) for three different values of p_0 . There is no way to tell, on meteorological or mathematical grounds, which is the correct temperature profile. A systematic variation can be seen, however. If the pressure is high, low temperatures are obtained at high levels and high temperatures at low levels. In between is a region in which the temperature is correct, irrespective of errors in p_0 .

The physical reasons for this behavior are easily understood [Gille and House, 1971]. Too high a pressure implies higher than correct densities everywhere, which increases the amount of emitting material and decreases transmittance. At upper levels, at which the optical depth to the tangent point d is less than 1, the increased emitting material dominates the decreased transmittance, requiring a lower than correct temperature for agreement with the observed radiance.

At levels where $d > 1$, the decreased transmittance dominates and requires a higher than correct temperature for agreement. The converse occurs for a pressure too low. At the level at which $d = 1$, the retrieved temperature is independent of p_0 .

This leads immediately to a method for determining p_0 (fig. 8). Two channels, with different spectral opacity and therefore with $d = 1$ at different levels, are employed. Each displays the behavior shown in figure 7, and in general the temperatures inferred in the two channels will not agree. If the pressure is too high, the opaque temperature will be higher than the transparent channel, and vice versa. Numerical studies [Gille and House, 1971] have shown that this temperature difference between the two channels is linear in pressure, with a zero crossing very close to p_0 . A consideration of all error sources suggests that p_0 may be determined to ± 0.3 mb, with ± 0.17 mb probably possible. Thus, successful determination of p_0 rests on accurate and rapid inversion of spectral data according to a scheme like the one outlined earlier, but for two spectral channels.

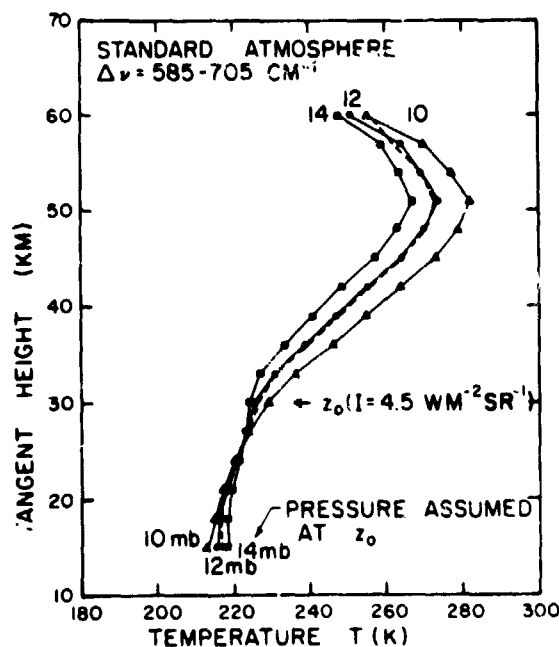


Figure 7. - The effect of the value of p_0 on the inferred profile. The dashed line is the true profile, the U.S. Standard Atmosphere. Solid lines are inversion solutions with the values 10, 12, and 14 mb assumed for p_0 . The correct value is 12 mb.

DETERMINATION OF TRACE CONSTITUENT AMOUNTS

If the temperature profile is known, measurements in suitable spectral channels may be inverted to obtain concentrations of variable constituents such as ozone and water vapor. Equation (1) is now the basic equation to invert, and the constituents are determined from their effects on the weighting function. This is a much more nearly linear problem, and one of the methods described earlier might be used. McKee et al., [1969] and Russell and Drayson [1972] have employed an "onion-peeling" method, working down from the top of the atmosphere to determine water vapor and ozone profiles from real and synthesized data. This is equivalent to solving a triangular matrix. House and Gille (in preparation) have employed an iterative method similar to the one described for temperature. Now the relaxation equation is

$$c^{n+1}(h^*) = c^n(h^*) \left[1 - \frac{1 - \tau^n(h^*)}{\tau^n(h^*) \ln \tau^n(h^*)} \frac{I(h^*) - I^n(h^*)}{I(h^*)} \right] \quad (7)$$

in which c is the mixing ratio. Usually no more than two iterations are necessary to obtain agreement within experimental error. Three examples of retrievals of realistic data at 3 km intervals to get stratospheric water vapor are shown in figure 9.

CONCLUSIONS

The results demonstrate that an iterative procedure of the type proposed by Chahine [1968] will enable temperature solutions to be retrieved from measurements of limb radiance in the CO_2 band, although the kernel is a strong, nonlinear function of the temperature. The algorithm also requires that the pressure at some level be specified; by starting with an initial guessed pressure and applying the algorithm to measurements in two spectral channels, the pressure may be accurately determined, and a self-contained inversion achieved. The inversion of trace constituent profiles is a nearly linear problem, and needs no special treatment.

ACKNOWLEDGMENTS

I thank Mr. Paul Bailey for running many of these calculations, and for useful discussions. This work was supported by the Atmospheric Sciences Section, National Science Foundation, under NSF Grant GA 20213 and by the Florida State University computing center.

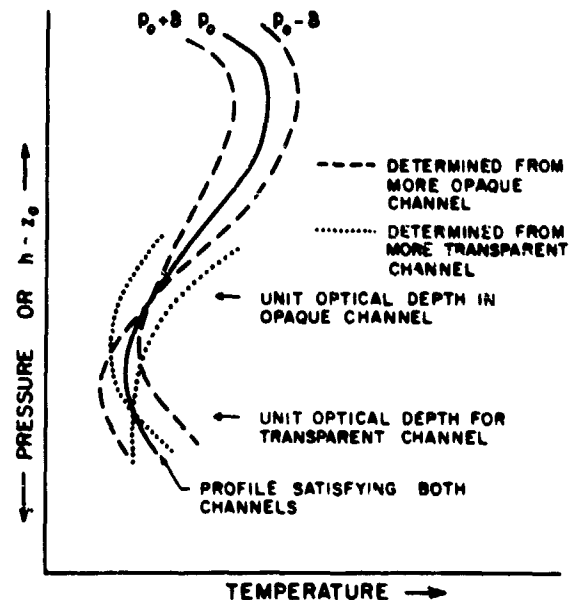


Figure 8. - Schematic diagram of the two-channel method for determining p_0 . Only the correct value results in a zero mean temperature difference between inversions of radiance data from the two channels.

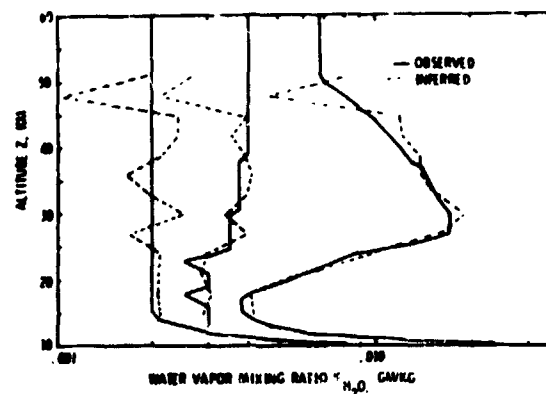


Figure 9. - Three inversions of synthesized realistic data to obtain distributions of water vapor in the stratosphere.

DISCUSSION

King: You indicated that noise is more of a limit to what you can infer or that noise prevents you from taking advantage of the edge on helping of your weighting function. Is that right?

Gille: Noise and the finite field of view are both factors, but noise at this level is more important.

PASSIVE ATMOSPHERIC SOUNDING

BIBLIOGRAPHY

- Alishouse, J. C., Crone, L. J., Fleming, H. E., Van Cleef, F. L., and Wark, D. Q., A Discussion of Empirical Orthogonal Functions and their Application to Vertical Temperature Profiles, Tellus, 19, 477-482, 1967.
- Baker, C. T. H., Fox, L., Mayers, D. F., Wright, K., Numerical Solution of Fredholm Integral Equations of First Kind, The Computer J., 7, 141-148, 1964.
- Bates, J. C., Hanson, D. S., House, F. B., Carpenter, R. O'B., and Gille, J. C., The Synthesis of 15μ Infrared Horizon Radiance Profiles from Meteorological Data Inputs, NASA, CR-724, 1967.
- Bjerhammar, A., Rectangular Reciprocal Matrices with Special Reference to Geodetic Calculations, Bull. Geodesique, 188-220, 1951.
- Bojkov, R. D., Differences in the Vertical Distribution Deduced From Umkehr and Ozonesonde Data at Goose Bay, J. Appl. Meteor., 5, 872-877, 1966.
- Catoe, G., Nordberg, W., Thaddeus, P., and Ling, G., Preliminary Results from Aircraft Flight Tests of an Electrically Scanning Microwave Radiometer, Goddard Space Flight Center, Tech. Rep. X-662-67-352, 1967.
- Chahine, M. T., Determination of the Temperature Profile in an Atmosphere From its Outgoing Radiance, J. Opt. Soc. Amer., 58, 1634-1637, 1968.
- Chahine, M. T., Inverse Problems in Radiative Transfer: Determination of Atmospheric Parameters, J. Atmos. Sci., 27, 960-967, Sept. 1970.
- Chahine, M. T., A General Relaxation Method for Inverse Solution of the Full Radiative Transfer Equation, 1272 to be published in J. Atmos. Sci.
- Chandrasekhar, S., Radiative Transfer, New York, Dover, 1960.
- Chaney, L. W., Drayson, S. R., and Young, C., Fourier Transform Spectrometer-Radiative Measurements and Temperature Inversion, Appl. Optics, 6, 347-349, 1967.
- Conrath, B. J., Inverse Problems in Radiative Transfer: A Review, Proc. XVIIIth Int. Astron. Cong., Academic Press, N.Y., 339-360, 1968.
- Conrath, B. J., On the Estimation of Relative Humidity Profiles from Medium-Resolution Infrared Spectra Obtained from a Satellite, J. Geophys. Res., 74, 3347-3361, 1969.
- Conrath, B. J., Hanel, R. A., Kunde, V. G., and Prabhakara, C., The Infrared Interferometer Experiment on Nimbus 3, J. Geophys. Res., 75, 5831-5857, 1970.
- Conrath, B. J., Hanel, R. A., Prabhakara, C., Kunde, V. G., Revah, I., and Salomonson, V. V., Vertical Sounding of the Atmosphere with Nimbus 4 Infrared Spectrometer Experiment, to be published in Proc. of 21st International Astronautical Congress, North Holland Publishing Co., Amsterdam, 1971.
- Craig, R. A., DeLuisi, J. J., and Steutger, I., Comparison of Chemiluminescent and Umkehr Observations of Ozone, J. Geophys. Res., 72, 1667-1671, 1967.

- Dahm, W. K., and Goulard, R., Molecular Radiation and its Application to Diagnostic Techniques, Proc. Conference at MSFC, Oct. 5-6, 1967.
- Dave, J. V., Sheppard, P. A., and Walshaw, C. D., Ozone Distribution and the Continuum From Observations in the Region of the 1043 cm^{-1} band, Quart. J. Roy. Meteorol. Soc., **89**, 307-318, 1963.
- Dave, J. V., Meaning of Successive Iteration of the Auxiliary Equation in the Theory of Radiative Transfer, Astronphys. J., **140**, 1292-1303, 1964.
- Dave, J. V., and Mateer, C. L., A Preliminary Study on the Possibility of Estimating Total Atmospheric Ozone From Satellite Measurements, J. Atmos. Sci., **24**, 414-427, 1967.
- DeLuisi, J. J., A Study of the Effect of Haze Upon Umkehr Measurements, Quart. J. Roy. Met. Soc., **95**, 181-187, 1969.
- DeLuisi, J. J., and Furukawa, P. M., An Explanation for the Worldwide Anomaly in the Concentration of Ozone Above 40 km, J. Atmos. Sci., **27**, 968-970, 1970.
- DeLuisi, J. J., and Mateer, C. L., On the Application of the Optimum Statistical Inversion Technique to the Evaluation of Umkehr Observations, J. Appl. Meteorol., **10**, 328-334, 1971.
- Drayson, S. R., Li, S. Y., and Young, C., "Atmospheric Absorption by Carbon Dioxide, Water Vapor and Oxygen, Final Report Contract Cwb 11376, U. of Michigan, College of Engineering Report No. 08183-2-F., 1968.
- Dütsch, H. U., Vertical Ozone Distribution From Umkehr Observations, Arch. Meteorol. Geophys. Biok. A., **11**, 240-251, 1959.
- Dütsch, H. U., Two Years of Regular Ozone Soundings Over Boulder, Colorado, NCAR-TN-10, National Center for Atmospheric Research, Boulder, Colo., 1966.
- Dütsch, H. U., Zü'lig, W., and Ling, Ch., Regular Ozone Observation at Thalwil, Switzerland, and at Boulder, Colorado, LAPETH-1, Laboratory for Atmospheric Physics, Swiss Federal Technical Institute, Zurich, 1970.
- Fisibach, F., A Satellite Method for Pressure and Temperature Below 24 km, Bull. Amer. Meteorol. Soc., **46**, 528-532, 1965.
- Fleming, H. E., and Smith, W. L., Inversion Techniques for Remote Sensing of Atmospheric Temperature Profiles, Proc. 5th Symp. Temperature, Its Measurement and Control in Science and Industry, Wash., D. C., June 21-24, 1971.
- Florance, E. T., Proc. Interdisciplinary Workshop on Vertical Radiometric Probing held at GCA Corporation, April 23-24, 1965, GCA-TR 65-12-N, June 1965.
- Florance, E. T., Mathematical Investigation of the Nonlinear Inversion Method, Scientific Rep. No. 4, Contract NAS5-3352, GCA-TR-66-8-N, GCA Technology Division, 1966.
- Florance, E. T., Error Analysis of Inversion Techniques, Final Report, Contract No. NAS5-3352, GCA-TR-67-8-N, GCA Technology Division, 1967.
- Florance, E. T., Error Analysis of the Generalized Nonlinear Inversion Method. Final Report, Contract No. NAS5-10393, GCA-TR-68-3-N, GCA Technology Division, 1968.

- Foster, M., An Application of the Wiener-Kolmogorov Smoothing Theory to Matrix Inversion, J. Soc. Indust. Appl. Math., **9**, 387-392, 1961.
- Franklin, J. N., Well-Posed Stochastic Extensions of Ill-Posed Linear Problems, Calif. Inst. Tech., Willis H. Booth Computing Center Programing Rep. 135, Aug., 1969.
- Gaut, N. E., Barrett, A. H., and Staelin, D. H., Results Obtained From the Inversion of Simulated Atmospheric Water Vapor Spectra, Quart. Prog. Rep. 85, Res. Lab. of Electron., M.I.T., p. 16, April 15, 1967.
- Gille, J. C., ed., Proc. Third Interdisciplinary Workshop on Inversion of Radiometric Measurements, Florida State Univ., Dept. of Meteorology Technical Report No. 68-2, 1968.
- Gille, J. C., On the Possibility of Estimating Diurnal Temperature Variation at the Stratopause From Horizon Radiance Measurements, J. Geophys. Res., **73**, 1863-1868, 1968.
- Gille, J. C., Inversion of Radiometric Measurements, Bull. Amer. Meteorol. Soc., **49**, 903-912, Sept. 1968.
- Gille, J. C., and House, F. B., On the Inversion of Limb Radiance Measurements I: Temperature and Thickness, J. Atmos. Sci., **28**, 1427-1442, Nov. 1971.
- Glasko, V. B., and Timofeyev, Yu. M., The Solution of the Thermal Sounding Problem for the Atmosphere Using the Regularization Method, Izv., Atmospheric and Oceanic Physics (Translation by the A.G.U.), **4**, (No. 3), 170, 1968.
- Global Atmospheric Research Programme (GARP), 1967. Available from the National Academy of Sciences.
- Goody, R. M., Atmospheric Radiation. I. Theoretical Basis. London, Oxford University Press, 1964.
- Götz, F. W. P., Zum Strahlungsklima des Spitzbergen Sommers, Gerl. Beitr. z. Geophys., **31**, 119-154, 1931.
- Götz, F. W. P., Meetham, A. R., and Dobson, G. M. B., The Vertical Distribution of Ozone in the Atmosphere, Proc. Roy. Soc. (London), Ser. A., **145**, 416-446, 1934.
- Green, A. E. S., Attenuation by Ozone and the Earth's Albedo in the Middle Ultraviolet, Appl. Opt., **3**, 203-208, 1964.
- Greville, T. N. E., The Pseudo-Inverse of a Rectangular or Singular Matrix and its Application to the Solution of Systems of Linear Equations, SIAM Rev., **1**, 38, 1959.
- Grimmer, M., The Space-Filtering of Monthly Surface Anomaly Data in Terms of Pattern, Using Empirical Orthogonal Functions, Quart. J. Roy. Meteorol. Soc., **89**, 395-408, 1963.
- Gruner, P., Dämmerungerscheinungen, in Handbuch der Geophysik, VIII, Physik der Atmosphäre I. Linke, F., and Moller, F., eds. Berlin, Borntraeger, 432-526, 1942.
- Hanel, R. A., Determination of Cloud Altitude from A Satellite, J. Geophys. Res., **66**, 1300, 1961.

- Hanel, R. A., Conrath, B. J., Hovis, W. A., Kunde, V., Lowman, P. D., Prabhakara, C., and Schlachman, B., Infrared Spectroscopy Experiment for Mariner Mars 1971, Icarus, 12, 48-62, 1970.
- Hanel, R. A., and Conrath, B. J., Interferometer Experiment on Nimbus 3: Preliminary Results, Science, 165, 1258-1260, Sept. 1969.
- Hanel, R. A., and Conrath, B. J., Thermal Emission Spectra of the Earth and Atmosphere Obtained from the Nimbus 4 Michelson Interferometer Experiment, Nature, 228, 143-145, 1970.
- Hanel, R. A., Schlachman B., Rogers D., Vanous D., Nimbus 4 Michelson Interferometer, Appl. Opt. 10, 1376-1382, June 1971.
- Hanson, R. J., A Numerical Method for Solving Fredholm Integral Equations of the First Kind using Singular Values, SIAM J. Numer. Anal., 8, 616-622, Sept. 1971.
- Hariharan, T. A., Polarization of Reflected Solar Radiation Over Land, Sea and Cloud Surfaces, Pure Appl. Geophys., 77, 151-157, 1969.
- Hariharan, T. A., and Stephens, W. L., Measurements of Polarization of Atmospheric and Surface Scattered Radiation in the Visible Region. Scientific Report No. 2, Contract NGR 05-007-041, Institute of Geophysics and Planetary Physics, U.C.L.A., 1968.
- Herman, B. M., and Yarger, D. N., Estimating the Vertical Atmospheric Ozone Distribution by Inverting the Radiative Transfer Equation for Pure Molecular Scattering, J. Atmos. Sci., 26, 153-162, Jan. 1969.
- Hilleary, D. T., Heacock, E. L., Morgan, W. A., Moore, R. H., Mangold, E. C., and Soules, S. D., Indirect Measurements of Atmospheric Temperature Profiles from Satellites: III. The Spectrometers and Experiments, Mon. Weather Rev., 94, 367-377, June 1966.
- Hilleary, D. T., Wark D. Q., James D. G., An Experimental Determination of the Atmospheric Temperature Profile by Indirect Means, Nature, 205, 489-491, 1965.
- House, F. B., and Gill, J. C., On the Inversion of Limb Radiance Measurements II: Ozone and Water Vapor, 1971 (to be published) in J. Geophys. Res.
- Iozanas, V. A., Determining the Vertical Ozone Distribution in the Upper Atmospheric Layers From Satellite Measurements of Ultraviolet Solar Radiation Scattered by the Earth's Atmosphere, Geomag. Aeron., 8, 403-407, 1968.
- Iozanas, V. A., Krasnopol'skiy, V. A., Kuznetsov, A. P., and Lebedinskiy, A. I., Studies of the Earth's Ozonosphere From Satellites, Izv. Atmos. Oceanic Phys., 5, 77-82, 1969.a.
- Iozanas, V. A., Krasnopol'skiy, V. A., Kuznetsov, A. P., and Lebedinskiy, A. I., An Investigation of the Planetary Ozone Distribution From Satellite Measurements of Ultraviolet Spectra, Izv. Atmos. Oceanic Phys., 5, 219-223, 1969.b.
- James, D. G., Indirect Measurements of Atmospheric Temperature Profiles From Satellites: IV Experiments with the Phase I Satellite Infrared Spectrometer, Mon. Weather Rev., 95, 457-462, July 1967.

- Kaplan, L. D., Inference of Atmospheric Structure From Remote Radiation Measurements, J. Opt. Soc. Amer. 49, 1004-1007, Oct. 1959.
- King, J. I. F., The Radiative Heat Transfer of Planet Earth, in Scientific Uses of Earth Satellites, J. A. Van Allen, ed., Univ. Michigan Press, Ann Arbor, 1956.
- King, J. I. F., Inversion by Slabs of Varying Thickness, J. Atmos. Sci. 21, 324-326, 1964.
- King, J. I. F., Error Analysis on Inversion Techniques, Interim Scientific Report No. 1, Contract No. NASS-3352, GCA-TR-64-16-N, GCA Technology Division, 1964.
- King, J. I. F., Meteorological Inferences from Radiance Measurements, Final Report, Contract No. Cwb-10833, GCA Technology Division, 1965.
- Krasnopol'skiy, V. A., The Ultraviolet Spectrum of Solar Radiation Reflected by the Terrestrial Atmosphere, and its Use in Determining the Total Content and Vertical Distribution of Atmospheric Ozone, Geomag. Aeron. 6, 236-242, 1966.
- Kreiss, W. T., Meteorological Observations with Passive Microwave Systems, Boeing Doc. D1-82-0692, Feb. 1968.
- Krueger, A. J., Rocket Measurements of Ozone Over Hawaii, Symposium sur L'ozone Atmospherique, 2-7 Septembre, 1968, Monaco, Centre National de la Recherche Scientifique, Paris, 225-229, 1969.
- Krueger, A. J., Heath, D. F., and Mateer, C. L., High Level Ozone Distribution From Nimbus 4 and Comparison with Simultaneous Rocket Data, Trans. Amer. Geophys. Un. 52, 226 (abstract M83), 1971.
- Landweber, L., An Iteration Formula for Fredholm Integral Equations of the First Kind, Am. J. Math. 73, 615-624, 1951.
- Lenoir, W. B., and Kochler, R. R., Jr., Inference of the Atmospheric Temperature Profile From Simulated Microwave Measurements From Space, Quart. Prog. Rep. Res. Lab. Electron., MIT, 86, 27, 1967.
- Lorentz, E. N., Empirical Orthogonal Functions and Statistical Weather Prediction, Sci. Rep. No. 1, Statistical Forecasting Project, Dept. of Meteorol., MIT, 1956.
- Mateer, C. L., A Study of the Information Content of Umkehr Observations, PhD. Thesis, Univ. Michigan, 1964.
- Mateer, C. L., and Dütsch, H. U., Uniform Evaluation of Umkehr Observations From the World Ozone Network, Part I, National Center for Atmospheric Research Boulder, Colo., 1964.
- Mateer, C. L., On the Information Content of Umkehr Observations, J. Atmos. Sci., 22, 370-381, 1965.
- McClatchey, R. A., The Use of the 4.3 Micron CO₂ Band to Sound the Temperature of a Planetary Atmosphere, in Proc. Symp. Electromagnetic Sensing of the Earth From Satellites, Ralph Zirkind, ed., Brooklyn Polytechnic Institute Press, 1965.

- McKee, T. B., Inference of Temperature and Water Vapor Structure in the Stratosphere From Limb Radiance Profiles, NASA TM X-1943, 1970.
- McKee, T. B., Whitman, Ruth I., and Davis, R. E., Infrared Horizon Profiles for Summer Conditions From Project Scanner, NASA TN D-4741, 1968.
- McKee, T. B., Whitman, Ruth I., and Lambiotte, J. J., Jr., A Technique to Infer Atmospheric Temperature from Horizon Radiance Profiles, NASA TN D-5068, 1969. a.
- McKee, T. B., Whitman, Ruth I., and Lambiotte, J. J., Jr., A Technique to Infer Atmospheric Water Vapor Mixing Ratio From Measured Horizon Radiance Profile, NASA TN D-5252, 1969b.
- Mitra, S. K., The Upper Atmosphere, The Asiatic Society, Calcutta, 1952.
- Moore, E. M., General Analysis., Mem. Amer. Philos. Soc. 1, p. 8 and ch. 3, 29, 1935.
- Obukhov, A. M., The Statistically Orthogonal Expansion of Empirical Functions. Izv. Seriya Geofizicheskay, Akad. Nauk. USSR No. 3, 432-439. (English translation by A.G.U. 288-291) Nov. 1960.
- Penrose, R. A., A Generalized Inverse for Matrices, Proc. Cambridge Philos. Soc., 51, 406-413, 1955.
- Petryshyn, W. V., On a General Iterative Method for the Approximate Solution of Linear Operator Equations, Math. Comp., 17, 1-10, 1963.
- Phillips, D. L., A Technique for the Numerical Solution of Certain Integral Equations of the First Kind, J. Assoc. Comp. Mach., 9, 84-97, 1962.
- Powell, M. J. D., An Efficient Method for Finding the Minimum of a Function of Several Variables Without Evaluating Derivatives, Computer J., 7, 155-162, 1964.
- Prabhakara, C., Feasibility of Determining Atmospheric Ozone From Outgoing Infrared Energy, Mon. Weather Rev., 97, 307-314, April 1969.
- Prabhakara, C., Conrath, B. J., Hanel, R. A., and Williamson, E. J., Remote Sensing of Atmosphere Ozone Using the 9.6μ Band, J. Atmos. Sci., 27, 689-697, July 1970.
- Rawcliffe, R. D., and Elliott, D. D., Latitude Distribution of Ozone at High Altitudes, Deduced From a Satellite Measurement of the Earth's Radiance at 2840A, J. Geophys. Res., 71, 5077-5089, 1966.
- Regener, E., and Regener, V. H., Aufnahme des ultravioletten Sonnenspektrums in der Stratosphäre und die verticale Ozonverteilung, Phys. Zeit., 35, 788-793, 1934.
- Rodgers, C. D., Satellite Infrared Radiometer, A Discussion of Inversion Methods, Memo. No. 66.13, Clarendon Laboratory, Oxford University, 1966.
- Rodgers, C. D., Remote Sounding of the Atmospheric Temperature Profile in the Presence of Cloud. Quart. J. Roy. Meteorol. Soc., 96, 654-666, 1970.
- Rodgers, C. D., Some Theoretical Aspects of Remote Sounding of the Earth's Atmosphere, J. Quant. Spectrosc. Radiat. Transfer, 11, 767-777, June 1971.

Rodgers, C. D., and Walshaw, C. D., The Computation of Infrared Cooling Rates in Planetary Atmospheres, Quart. J. Roy. Meteorol. Soc., 92, 67-92, 1966.

Rosenkranz, P. W., Preliminary Results From the 1970 Airborne Meteorology Expedition, Quart. Prog. Rep. Res. Lab. Electron., MIT, 99, 1970.

Rozenberg, G. V., Twilight, A Study in Atmospheric Optics, New York, Plenum Press. (translated from Russian), 1966.

Russell, F. A. R., and Archibald, E. D., On the Unusual Optical Phenomena of the Atmosphere, 1883-1886, Including Twilight Effects, Coronal Appearances, Sky Haze, Colored Suns, Moons, etc., in The Eruption of Krakatoa, and Subsequent Phenomena, G. J. Symons, ed., 151-463, Roy. Soc., London, 1888.

Russell, J. M., III, The Measurement of Atmospheric Ozone Using Satellite Infrared Observations in the $9.6 \mu\text{m}$ Band. PhD. Thesis, University of Michigan, 1970 (also College of Engineering Report 3635-1-T).

Russell, J. M., III, and Drayson, S. R., The Inference of Atmospheric Ozone Using Satellite Horizon Measurements in the 1042 cm^{-1} band, J. Atmos. Sci., 28, 1971.

Saiedy, F., Hilleary, D. T., and Morgan, W. A., Cloud-top Altitude Measurements From Satellites, Appl. Opt., 4, 495-500, 1965.

Scott, N. A., and Chedin, A., A Least-Square Procedure Applied to the Determination of Atmospheric Temperature Profiles from Outgoing Radiance, J. Quant. Spectrosc. Radiat. Transfer, 11, 405-420, 1969.

Sekera, Z., Determination of Atmospheric Parameters From Measurement of Polarization of Upward Radiation by Satellite or Space Probe, Icarus, 6, 348-359, 1967.

Sellers, W. D., and Yarger, D. N., The Statistical Prediction of the Vertical Ozone Distribution, J. Appl. Meteorol., 8, 357-361, 1969.

Shaw, C. B., Jr., Improvement of the Resolution of an Instrument by Numerical Solution of an Integral Equation, Autonetics Rep. X70-510/501 (Autonetics, A Division of North American Rockwell Corp., 3370 Miraloma, Anaheim, Calif. 92803), 1969.

Shaw, F. S., Relaxation Methods, Dover Publications, Inc., New York, 1953.

Shen, W. C., and Smith, W. L., On the Discrepancy Between Calculated and Observed Nimbus II $6.7 \mu\text{m}$ Water Vapor Radiation, J. Appl. Meteorol., 10, 575-581, 1971.

Singer, S. F., and Wentworth, R. C., A Method for the Determination of the Vertical Ozone Distribution From a Satellite, J. Geophys. Res., 62, 299-308, 1957.

Smith, W. L., An Iterative Method for Deducing Tropospheric Temperature and Moisture Profiles From Satellite Radiation Measurements, Mon. Weather Rev., 95, 363-369, 1967.

- Smith, W. L., An Improved Method for Calculating Tropospheric Temperature and Moisture From Satellite Radiometer Measurements, Mon. Weather Rev., **96**, 387-396, June 1968.
- Smith, W. L., Statistical Estimation of the Atmosphere's Geopotential Height Distribution From Satellite Radiation Measurements, ESSA Technical Repor. NESC 48, National Environment Satellite Center, 1969.
- Smith, W. L., Iterative Solution of the Radiative Transfer Equation for the Temperature and Absorbing Gas Profile of an Atmosphere, Appl. Opt., **9**, 1993-1999, 1970.
- Smith, W. L., and Howell, H. B., Vertical Distribution of Atmospheric Water Vapor From Satellite Infrared Spectrometer Measurements, Presented at the 51st Annual Meeting of the Amer. Meteorol. Soc., San Francisco, Calif., 12 Jan. 1971, J. Appl. Meteorol. **10**, 1026-1034, after Aug. 1971.
- Smith, W. L., and Wark, D. Q., Atmospheric Soundings Derived From Nimbus III Spectrometer Measurements, Proc. Tech. Conf. Upper Air Instruments and Observations, World Meteorological Organization, Paris, France, Sept. 1969.
- Smith, W. L., Woolf, H. M., and Jacob, W. J., A Regression Method for Obtaining Real Time Temperature and Geopotential Height Profiles From Satellite Spectrometer Measurements and its Application to Nimbus III "SIRS" Observations, Mon. Weather Rev., **98**, 582-603, 1970.
- Smith, W. L., Woolf, H. M., and Fleming, H. E., Retrieval of Atmospheric Temperature Profiles From Satellite Measurements for Dynamical Forecasting, J. Appl. Meteorol., to be published in Feb 1972.
- Scholev, V. V., A Treatise on Radiative Transfer, Princeton, Van Nostrand, (translated from Russian), 1963.
- Special Issue on Remote Environment Sensing, Proc. IEEE, April 1969
- Staelin, D. H., Measurements and Interpretation of the Microwave Spectrum of the Terrestrial Atmosphere near 1-centimeter Wavelength, J. Geophys. Res., **71**, 2875-2881, June 15, 1966.
- Staelin, D. H., Interpretation of Spectral Data, on p. 15 in Quart. Prog. Rep. 85, Res. Lab. Electron., M.I.T., April 15, 1967.
- Staelin, D. H., Passive Remote Sensing at Microwave Wavelengths, Proc. I.E.E.E., **57**, 427-439, April 1969.
- Strand, O. N., and Westwater, E. R., Statistical Estimation of the Numerical Solution of a Fredholm Integral Equation of the First Kind, J. Assoc. Comp. Mach., **15**, 100-114, 1968.a.
- Strand, O. N., and Westwater, E. R., Minimum RMS Estimation of the Numerical Solution of a Fredholm Integral Equation of the First Kind, SIAM J. Numer. Anal., **5**, 287, 1968.b.
- Strand, O. N., Optimization of Frequencies Used in Indirect Sensing by Inversion, NOAA Tech. Rpt. ERL202-WPL14, U.S. Dept. Of Commerce, NOAA-ERL, March 1971.
- Taylor, F. W., Temperature Sounding Experiments for the Jovian Planets, to be published in J. Atmos. Sci., 1972.

- Turchin, V. F., Malkevich, M. S., and Gorchakova, I. A., The Use of Statistical Regularization Determining the Vertical Atmospheric Temperature Profile, Bull. (Izv.) Acad. Sci. USSR Atm. and Oceanic Phys., 5, 252-256 (English translation) 1969.
- Turchin, V. F., and Nozik, V. Z., Statistical Regularization of the Solution of Incorrectly Posed Problems, Bull. (Izv.) Acad. Sci. USSR Atm. and Oceanic Phys., 5, 14-18. (English translation) 1969.
- Twomey, S., On the Deduction of the Vertical Distribution of Ozone by Ultraviolet Spectral Measurements From a Satellite, J. Geophys. Res., 66, 2153-2162, 1961.
- Twomey, S., On the Numerical Solution of Fredholm Integral Equations of the First Kind by the Inversion of the Linear System Produced by Quadrature, J. Assoc. Comp. Mach., 10, 97-101, 1963.
- Twomey, S., and Howell, H. B., A Discussion of Indirect Sounding Methods with Special Reference to the Deduction of Vertical Ozone Distribution From Light Scattering Measurement, Mon. Weather Rev., 91, 659-664, 1963.
- Twomey, S., The Application of Numerical Filtering to the Solution of Integral Equations Encountered in Indirect Sensing Measurements, J. Franklin Inst., 279, 95-109, 1965.
- Twomey, S., Indirect Measurements of Atmospheric Temperature Profiles From Satellites. II. Mathematical Aspects of the Inversion Problem, Mon. Weather Rev., 94, 363-366, 1966.
- Twomey, S., and Howell, H. B., Some Aspects of the Optical Estimation of Microstructure in Fog and Cloud, Appl. Opt., 6, 2125-2131, 1967.
- Twomey, S., Comments on Estimating the Vertical Ozone Distribution, J. Atmos. Sci., 26, 1154, 1969.
- Wahba, G., On the Numerical Solution of Fredholm Integral Equations of the First Kind, U. of Wisconsin, Dept. Statistics, Tech. Rep. 217, Oct., 1969.
- Wahba, G., Convergence Rates for Certain Approximate Solutions to Fredholm Integral Equations of the First Kind, Dept. Statistics, U. Wisconsin, 1970.
- Wahba, G., A class of Approximate Solutions to Linear Operator Equations, Dept. of Statistics, U. Wisconsin, Tech. Rep. 270, July 1971.
- Wang, J. Y., Theory and Applications of Inversion Techniques: A Review, Report AA & ES 70-6, Purdue University School of Aeronautics, Astronautics and Engineering Sciences, 1970.
- Wark, D. Q., On Indirect Soundings of the Stratosphere From Satellites, J. Geophys. Res., 66, 77-82, 1961.
- Wark, D. Q., SIRS: An Experiment to Measure the Free Air Temperature From a Satellite, Appl. Opt., 9, 1761-1766, 1970.
- Wark, D. Q., and Fleming, H. E., Indirect Measurements of Atmospheric Temperature Profiles From Satellites. I. Introduction, Mon. Weather Rev., 94, 351-362, 1966.

Wark, D. Q., and Hilleary, D. T., Atmospheric Temperature: Successful Test of Remote Probing Science, 165, 1256-1258, Sept. 1969.

Wark, D. Q., Hilleary, D. T., Fleming, H. E., Smith, W. L., and Lienesch, J. H., Atmospheric Temperature Determinations From the SIRS-A on Nimbus III, Proc. Sixth International Symposium on Remote Sensing of Environment, Univ. Michigan, Oct. 1969.

Wark, D. Q., and Mercer, D. M., Absorption in the Atmosphere By the Oxygen 'A' Band, Appl. Opt., 5, 1469, 1965.

Wark, D. Q., Saiedy, F., and James, D. G., Indirect Measurements of Atmospheric Temperature Profiles From Satellites: VI. High Altitude Balloon Testing, Mon. Weather Rev., 95, 468-479, 1967.

Waters, J. W., and Staelin, D. H., Statistical Inversion of Radiometric Data, Quart. Prog. Rep. 89, Res. Lab. Electron., M.I.T., April 15, 1968.

Waters, J. W., and Staelin, D. H., Results of Inferring Atmospheric Temperature From Simulated Microwave Measurements, Res. Labs. Electron. MIT QPR 90, 11, 1968.

Westwater, E. R., Ground-Based Determination of Temperature Profiles by Microwaves, Univ. of Colorado, Ph.D. Thesis, 1970. (Also Mon. Weath. Rev., 100, Jan. 1972.)

Westwater, E. R., and Strand, O. N., Application of Statistical Estimation Techniques to Ground Based Passive Probing of the Tropospheric Temperature Structure, Environmental Sciences Services Administration Tech. Rep. IER37-ITSA37, 1967.

Westwater, E. R., and Strand, O. N., Statistical Information Content of Radiation Measurements Used in Indirect Sensing, J. Atmos. Sci., 25, 750-758, Sept. 1968.

Whitman, Ruth I., Effects of Three Transmission Models in the Rotational Water Vapor Band on Radiance Calculations and Constituent Inferences, NASA TN D-6112, 1971.

Whitman, Ruth I., McKee, T. B., and Davis, R. E., Infrared Horizon Profiles for Winter Conditions From Project Scanner, NASA TN D-4905, 1968.

Wolk, M., VanCleaf, F., and Yamamoto, G., Indirect Measurements of Atmospheric Temperature Profiles From Satellites: V. Atmospheric Soundings From Infrared Spectrometer Measurements at the Ground, Mon. Weather Rev., 95, 463-467, July 1967.

Yamamoto, G., Numerical Method for Estimating the Stratospheric Temperature Distribution From Satellite Measurements in the CO₂ Band, J. Meteorol., 18, 581, 1961.

Yamamoto, G., and Wark, D. Q., Discussion of the Letter by R. A. Hanel, 'Determination of Cloud Altitude From a Satellite', J. Geophys. Res., 66, 3596, 1961.

Yarger, D. N., An Evaluation of Some Methods of Estimating the Vertical Atmospheric Ozone Distribution From the Inversion of Spectral Ultraviolet Radiation, J. Appl. Meteorol., 9, 921-928, 1970.

2. ACTIVE ATMOSPHERIC SOUNDING

In contrast to the passive (radiance) measurement techniques discussed in Chapter 1, this chapter is concerned with measurements of reflected and/or scattered radiation from the atmosphere arising from man-made probing sources—that is, active radar experiments. Both monostatic and bistatic configurations with pulse and continuous-wave techniques are in use, as well as acoustic and electromagnetic (radio-frequency and optical (lidar)) radars.

The data are inverted to yield vertical profiles of wind speed and direction, turbulence, precipitation, refractive index structure, particulate concentrations, and gaseous molecule concentrations. Active remote sounding is a rapidly growing research field, and great technological advances leading to sophisticated sounding techniques have been made in recent years; however, because this is a relatively new field, the inversion algorithms that have been developed specifically for this discipline are still rather primitive. It is clear that scientists active in this area should benefit enormously from algorithms developed and described elsewhere in this volume.

A. M. Peterson organized and chaired the session devoted to active atmospheric sounding.

ACOUSTIC RADAR SOUNDING OF THE LOWER ATMOSPHERE

L. G. McAllister

Australian Defence Scientific Service
Department of Supply
Weapons Research Establishment
Salisbury, South Australia

N73-11592 ABSTRACT

Acoustic radar sounding techniques were used to measure the wind velocity and direction in the first 300 m of the atmosphere. Angle-of-arrival and Doppler techniques were developed to obtain two independent measurements of the wind field. These techniques and preliminary experimental results are described briefly.

INTRODUCTION

The potential of the acoustic radar sounding technique for the remote measurement of the structure of the lower atmosphere (wind velocity, temperature, and so forth), has been recognized for some time [McAllister, 1968; McAllister et al., 1969; McAllister and Pollard, 1969; Little, 1969; Beran, 1970]. This technique consists of transmitting short pulses (40-200 msec) of acoustic energy at a carrier frequency of 800 Hz at a power level of 100 W from an acoustic array on the ground, and recording the energy backscattered from turbulent fluctuations of temperature and of wind velocity on a facsimile recorder, to give a height versus time of day display of the structure of turbulence in the first few kilometers of the atmosphere. It has been shown [McAllister, 1968; McAllister, et al. 1969; McAllister and Pollard, 1969] that thermal plumes, radiation inversions, and sea breeze circulations can be studied with the aid of the acoustic sounder. More recently, a joint acoustic sounding experiment was conducted with the Wave Propagation Laboratory (WPL) of the National Oceanic and Atmospheric Administration (NOAA) Boulder, Colorado, when the sounder was operated alongside a 150 m meteorological tower situated at Haswell, Colorado. Good progress was made toward interpreting the vast amount of detail contained in the sounder record in terms of the profiles of temperature and wind velocity, and the turbulent flux of heat and momentum [Bean and McAllister, 1971].

During this series of experiments, however, there were some problems in the use of tower data to interpret the sounding data. First, it was necessary to separate the tower and the sounder by 300 m to avoid echoes from the tower structure; hence the correlation between tower and sounder data was not always good during nonstationary, atmospheric processes. Second, the levels of instrumentation on the tower were limited by practical considerations to 40 m, 95 m, and 150 m; frequently, the height range of interest lay between these levels and no tower data could be obtained for this range. Third, the sounder was a more sensitive indicator of temperature fluctuations than were the temperature-sensing elements on the tower, and it was not always possible to investigate the structure recorded by the sounder under very stable conditions in light winds. The need to develop the acoustic sounder into an independent, quantitative, meteorological instrument for interpretation of fine detail recorded during nonstationary processes was thus highlighted.

Angle of arrival [McAllister et al., 1969] and Doppler [Kelton and Bricout, 1964] techniques have been added to the sounder to obtain two independent measurements of wind velocity. Here we discuss these techniques and present the preliminary results of the experiments.

LIST OF SYMBOLS

$\sigma(\theta)$	scattered power, per unit volume, per unit incident flux, per unit solid angle at an angle θ from the initial direction of propagation
λ	acoustic wavelength
k	wave number of the sound wave
C_v	structure constant of the wind
C_t	structure constant of the temperature
C	velocity of sound
T	absolute temperature
w	horizontal wind velocity
V_a	total wind velocity measured by anemometer
V_θ	total wind velocity measured by angle of arrival technique
V_r	total wind velocity measured by Doppler technique
A_z^a	wind direction at height of anemometer
A_z^θ	wind direction measured by angle of arrival technique
A_z^r	wind direction measured by Doppler technique

SCATTERING OF SOUND IN A TURBULENT ATMOSPHERE

The propagation of sound through a turbulent medium has been studied for some time, and the work of Ford and Meeham [1960], Tataraki [1961], Kallistratova [1961], and Monin [1962] has resulted in an equation describing the scatter of sound waves by turbulent fluctuations of temperature and of wind velocity. The equation is:

$$\sigma(\theta) = 0.03k^{1/3} \cos^2 \theta \left[\frac{C_v^2}{C^2} \cos^2 \frac{\theta}{2} + 0.13 \frac{C_t^2}{T^2} \right] \left(\sin \frac{\theta}{2} \right)^{-\frac{11}{3}} \quad (1)$$

where σ is the scattered power, per unit volume, per unit incident flux, per unit solid angle at an angle θ from the initial direction of propagation; k is the wave number of the sound wave; and C_v and C_t are the structure constants of wind and of temperature and are functions of the root mean square difference in longitudinal velocity (or in temperature) for two points separated by unit distance.

Equation (1) shows that the scattered power is the sum of two terms, one for wind fluctuations and the other for temperature fluctuations. The wind term has a multiplying factor of $\cos^2 (\theta/2)$, which means that the wind fluctuations produce no energy in the backscatter direction that is, for $\theta = 180^\circ$. Hence, all the energy received in the backscattered direction results from the interaction of the sound wave with turbulent fluctuations of temperature. At all other angles, the scattered energy is the sum of both the wind and temperature terms.

MEASUREMENT OF MEAN WIND VELOCITY

The vertically directed beam from the acoustic sounder is transported horizontally by the wind because the sound energy travels as a spherical wave relative to the air stream. This property of the propagation of sound is used to measure the mean velocity of the wind up to the height of the scattering volume.

In the presence of a constant horizontal wind w , a sound ray transmitted upwind at an angle θ to the vertical direction given by

$$\theta = \arcsin \frac{w}{C} \quad (2)$$

(where C is the velocity of sound) will reach the scattering layer, vertically above the transmission point. After scattering, the sound ray drifts horizontally on the return journey and arrives back at the receiving point from the downwind direction at an angle θ . Thus the angular difference between the transmitted and received rays is 2θ in the presence of a constant wind. If the horizontal wind is not constant along the whole path, the angular difference between the two rays becomes a function of the integral of the wind velocity along the path.

MEASUREMENT OF ANGLE OF ARRIVAL

The angle of arrival of the backscattered energy was measured using interferometry techniques. Four acoustic arrays were used in this series of experiments; the arrays are shown mounted on a horizontal turntable in figure 1.

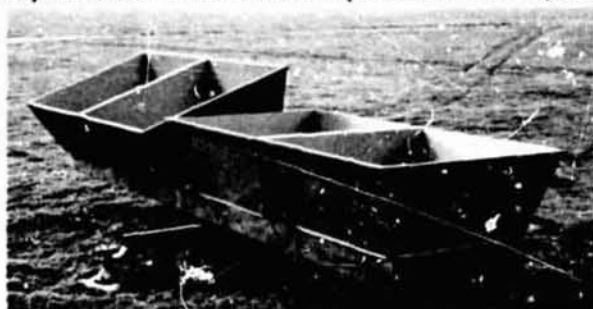


Figure 1. - Two oblique arrays and two vertical arrays mounted on a horizontal turntable.

Each array consisted of 49 loudspeakers with a diameter of 20 cm arranged in a 7×7 square array. The spacing between the rows of loudspeakers was 22 cm. Wooden side screens were fitted to each array to reduce the sidelobe level in the horizontal direction. When the arrays were grouped in pairs as shown, the distance between the phase centers of the arrays was 2.14 m (5λ).

Two receiving arrays were placed side by side, and the phase difference between the energy received in the two arrays was measured in a phase detector.

To minimize errors introduced by receiver noise, localized ambient noise sources, and other factors, the mean phase difference between these arrays was kept at zero by electrically steering the transmitted beam with a servo loop, which compensated for changes introduced by the wind. The phase difference between the two transmitting arrays that was required to steer the beam was a measure of the mean velocity of the wind.

In the present case, the transmitted energy was contained within a cone defined by the beam width of the acoustic array; hence the backscattered energy, which arrived back at the receiving array, had a similar spread in angle of arrival about the boresight direction. The mean direction of arrival was obtained by averaging the angle-of-arrival measurements over a period of several minutes. Range gating techniques were used to define the height range of the measurement.

The acoustic sounder was operated in the bistatic mode (figure 2(c)) to check the accuracy of the angle of arrival measurement. The pair of inclined arrays on the left of figure 1 was mounted over the vertical axis of the turntable and inclined at 20° to the horizontal as shown. The second pair of arrays was mounted on a fixed trestle, 100 m away toward the west and inclined at 20° to the horizontal, facing the pair on the turntable as shown in figure 2(c). This symmetrical, bistatic configuration gave a scattering angle of 140° , and the scattered power therefore contained components arising from fluctuations of temperature and of wind velocity (Eq. (1)) within the common volume of the two overlapping beams. The mean height of the common volume was 137.5 m.

The turntable was rotated in 3° increments through a total angle of 15° . Figure 3 is a plot of the turntable azimuth direction versus the phase change introduced by the servo loop to steer the transmitted beam back to the

original direction. Each measurement took 10 min; the extreme values of the phase angle recorded during this time are plotted in figure 3, with theoretical curve shown (solid line) for comparison.

EXPERIMENTAL MEASUREMENT OF MEAN WIND VELOCITY

The four acoustic arrays were arranged as shown in figure 2. The north-east and north-west arrays were used to measure the west-east component of the wind, and the north-west and south-west arrays were used to measure the north-south component of the wind. The south-east array was used to record the amplitude of the backscattered energy. The two components of the wind velocity were measured by time sharing the measuring equipment between the north pair of arrays and the west pair of arrays. Each pair was alternately switched into the measuring equipment before successive transmitted pulses. In this way it was possible to obtain continuous recordings of the northerly and easterly components of the wind.

On the night of April 2, 1971, a wind shear layer developed within the radiation inversion and the structure recorded by the sounder was similar to that shown in figure 10. The upper limit of turbulent breakdown (top of the wind shear layer) was at 180 m. The range gate was set between 90 and 150 m. The top trace in figure 4(a) is the total mean wind velocity constructed from the northerly and easterly components recorded by the sounder. This is the average wind velocity up to a height of 120 m. The dotted trace is the wind velocity recorded by an anemometer at a height of 10 m. Note the excellent agreement between the overall shape of the two traces and the higher velocity recorded by the sounder. Good correlation between these two traces would be expected because the mean velocity recorded by the sounder includes the velocity of the wind at the surface. In figure 4(b), the solid trace shows the direction of the wind recorded by the sounder, and the dotted trace shows the direction recorded by the anemometer. Unfortunately, there were only eight segments on the wind direction indicator mounted on the anemometer, and good angular resolution was not possible; but again, the similarity in the two traces is evident. Significant changes in the direction of the wind with height would not be expected through a wind shear layer of this kind. Therefore, the ability to record faithfully the direction of the wind under these conditions leads to confidence in the angle-of-arrival measurement.

In figure 4(c), the total wind velocity recorded by the sounder (V_θ) is plotted against the wind velocity recorded by the anemometer (V_A). Each dot represents a 1.5-min. sample. The good correlation between V_θ and V_A is seen clearly.

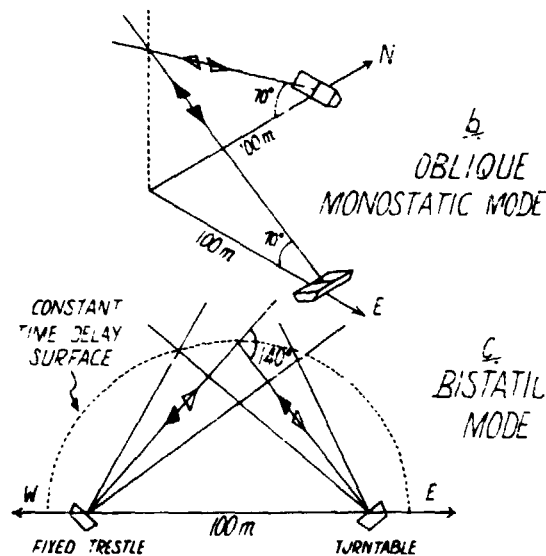
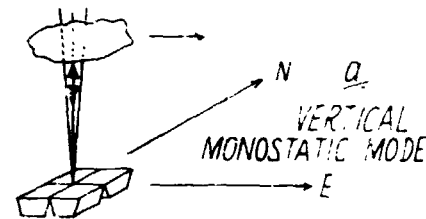


Figure 2.—Diagram showing modes of operation of acoustic sounder, (a) Vertical monostatic mode, (b) Oblique monostatic mode, (c) Bistatic mode.

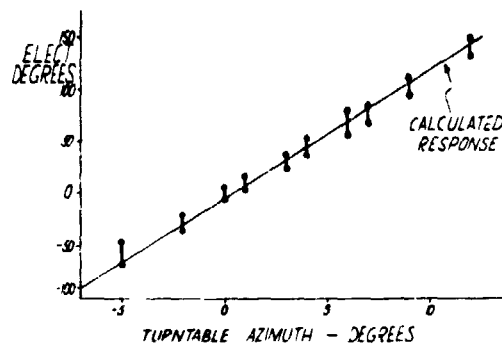


Figure 3.—Curve showing response of angle of arrival feedback loop to rotation of turntable in azimuth. The full line shows the calculated response.

DOPPLER MEASUREMENT OF WIND VELOCITY

The Doppler spectrum of an acoustic signal scattered from isotropic turbulence has been discussed theoretically by *Ford and Meecham* [1960] and investigated experimentally by *Kelton and Bricout* [1964]. *Ford and Meecham* showed that the maximum of the scattered power spectrum is shifted from the incident frequency by an amount determined by the Doppler shift due to the radial component of the mean flow.

The shift in the spectrum of the scattered power was measured by passing the received signal through a first-order tracking filter in which the capture range was set at ± 15 Hz. This gave a time constant of 65 msec for the tracking filter that was adequate to track a transmitted pulse of 80-msec duration. The output of the tracking filter was a voltage that was proportional to the frequency shift of the scattered power spectrum relative to the transmitted frequency. Null detection techniques were used again (see angle-of-arrival measurements) to reduce the errors introduced by receiver noise and narrowband, ambient noise sources. The output of the tracking filter was used to control the carrier frequency of the transmitted pulse in such a way that an increase in the received frequency due to the Doppler effect was compensated by a decrease in carrier frequency. Sufficient gain was applied around the servo loop to keep the output of the tracking filter near zero. The time constant of the Doppler servo loop was set at 30 sec.

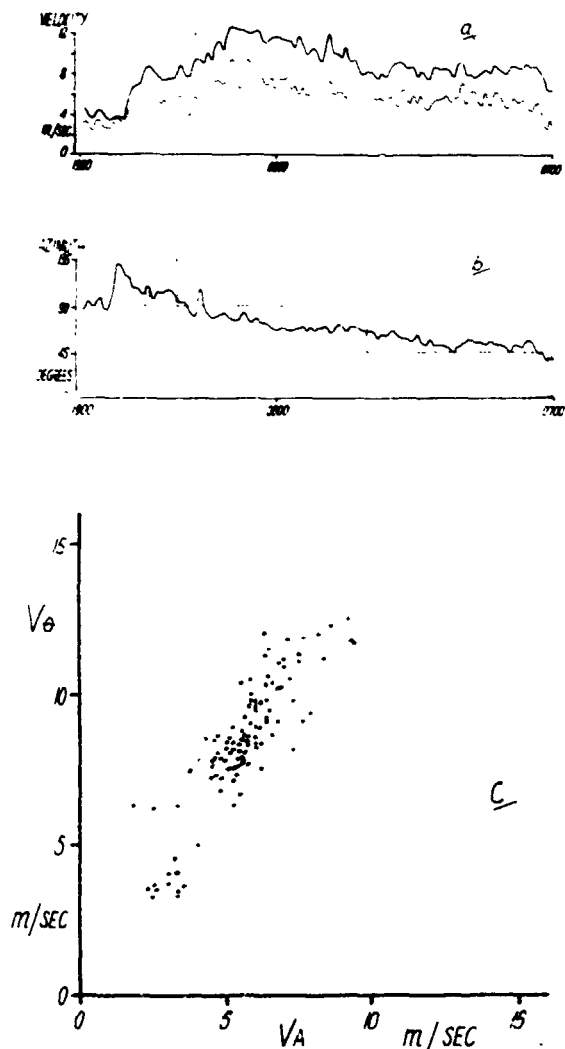


Figure 4.— Comparison between total wind velocity and direction measured by angle of arrival technique (full line) and anemometer (dotted line) at a height of 10 m.

(a) Total wind velocity.

(b) Direction, 1900 C.S.T. April 1, 1971 to 0700 C.S.T. April 2, 1971.

(c) Correlation between total wind velocity measured using angle of arrival technique (V_θ) and anemometer (V_A).

facsimile record. The top trace in figure 5 is the vertical component of the wind, drawn to the same time scale as the facsimile record. Note the correspondence between the columns of intense acoustic returns and periods of upward motion measured using the Doppler technique. Similarly, downward motion is seen to be associated with the relatively clear regions between the thermal columns. The peak vertical velocity recorded in this sample, was of the order of 1.6 m/sec.

The facsimile record shown in figure 6 was obtained using the symmetrical, bistatic configuration of the sounder (figure 2(c)) during thermal plume activity. It will be noticed that the height scale is now nonlinear, since time delay (linear scale on facsimile record) and height are not linearly related in the bistatic mode. The ellipsoidal surface of constant time delay, which passes through the center of the common volume, is drawn schematically in figure 2(c).

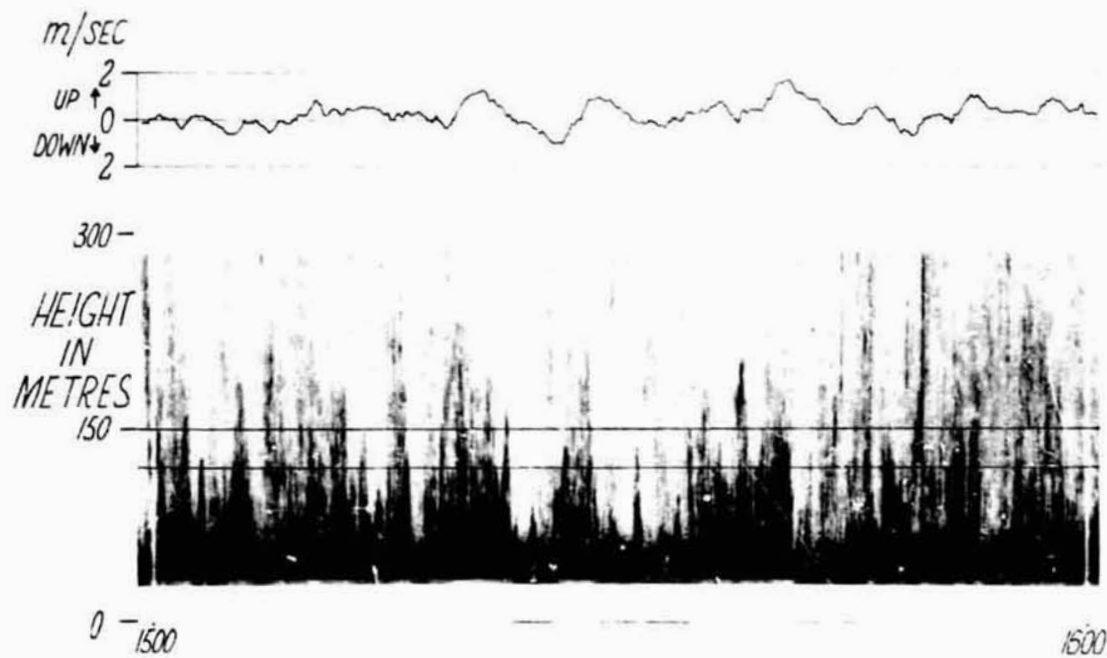


Figure 5.— Top Trace — Doppler measurement of vertical wind velocity.
Bottom Trace — Simultaneous vertical monostatic acoustic sounder
record of thermal plumes, 1500-1600 C.S.T. April 2, 1971. Height
of range gate shown by parallel lines on record.

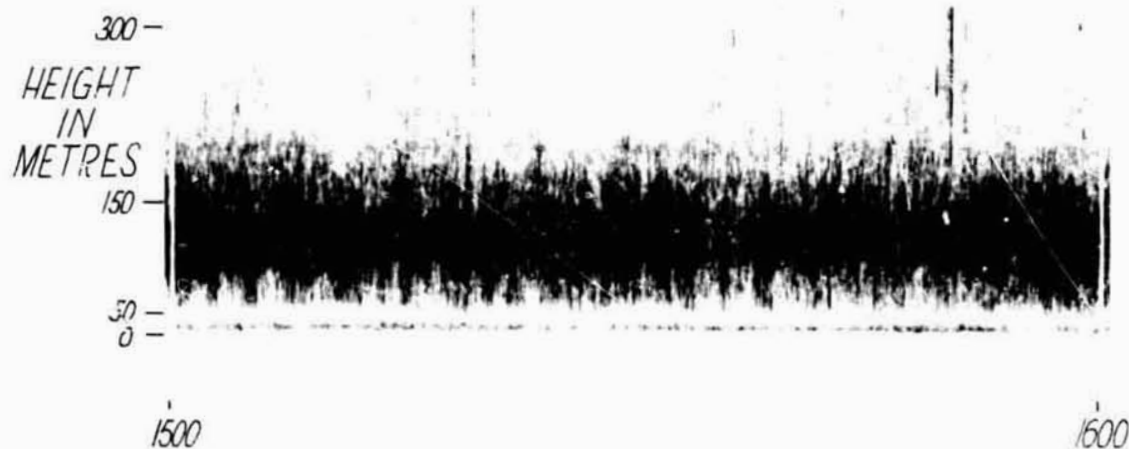


Figure 6.— Bistatic acoustic sounder record of thermal plumes 1500-1600 C.S.T. April 7, 1971.

The contrast between the records of figures 5 and 6, which are records of thermal plume activity, is striking. In figure 6, the clear regions between the thermal plumes (fig. 5) are now filled in with returns and the detail is lost. The additional energy was scattered from turbulent fluctuations of wind velocity since, in the bistatic mode of operation, the scatter angle was 140° (Eq. (1)). The clear regions between the thermal plumes shown in figure 5 (monostatic mode) indicate that the fluctuations of temperature were small in the descending air [Warner and Telford, 1967]. The apparent fluctuations of wind velocity in this descending air, which were recorded in the bistatic mode (fig. 6), probably arose from mechanically induced turbulence.

The difficulties in using other meteorological instrumentation to interpret the qualitative (and now quantitative) data recorded by the sounder and the need to develop the sounder into an independent instrument have been noted. The approach has been to (1) devise independent techniques to measure each parameter simultaneously and (2) obtain at least two independent streams of data that can be cross checked for error with the hope that confidence eventually will be built up in the results of the measurements through self consistency. These principles are illustrated in the following sections.

INDEPENDENT DOPPLER MEASUREMENTS OF VERTICAL VELOCITY

The vertical component of the wind velocity was measured using the symmetrical bistatic mode of sounder operation (fig. 2(c)). Independent streams of data were obtained by reversing the direction of transmission before each transmitted pulse. The west array, for example, was alternately switched to the transmitter and then to the receiver. The interleaved data so obtained were separated and recorded in separate channels. The upper two traces in figure 7 are the two channels of vertical wind velocity. The third trace from the top is the difference between the first two and was recorded at the same sensitivity. The good correlation between the two channels of vertical velocity is reflected in the difference channel. There was very little change in the character of the difference channel from 1300 to 1700 C.S.T., April 7, 1971, even though there were significant changes in the velocity channels. The fourth and fifth traces are the wind velocity and temperature measured at 10 m from the ground. The record was obtained after a sea breeze had developed at the experimental site (10 km from the coast). The temperature trace

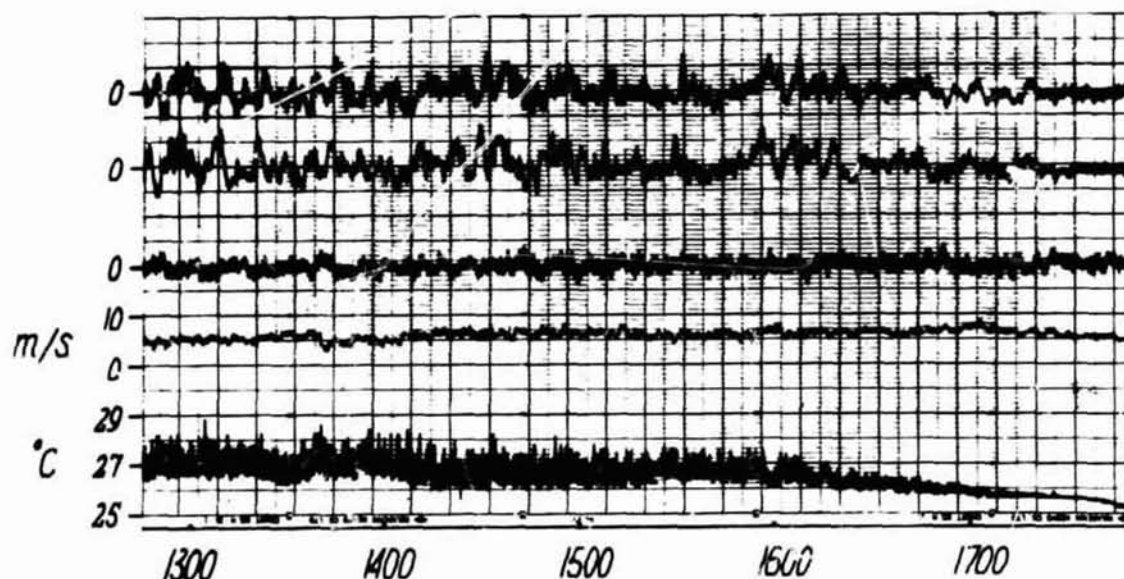


Figure 7.— Correlation between independent Doppler measurements of vertical wind velocity (V_p). Traces 1 and 2 (from top), V_{p1} and V_{p2} ; Trace 3, V_{p1} minus V_{p2} ; Trace 4, anemometer wind velocity (V_a); Trace 5, Temperature (10 m). 1300-1700 C.S.T. April 7, 1971.

clearly shows the decline of surface heating and thermal plume activity after 1600, which is reflected in the velocity traces as a decline in vertical wind velocity after 1630. The mean value of the vertical wind between 1300 and 1500 was zero; that is, there was no net upward or downward flow of air that was measurable. The rms velocity during this period was 0.51 m/sec with peaks ranging between +1.3 m/sec and -1.2 m/sec.

SIMULTANEOUS MEASUREMENT OF THE HORIZONTAL WIND BY ANGLE-OF-ARRIVAL AND DOPPLER TECHNIQUES

The oblique monostatic mode of sounder operation was used to obtain the horizontal wind vector by the two independent techniques of angle of arrival and Doppler. One pair of arrays was placed at right angles to the north-south arm (fig. 2(b)) and the boresight direction was set at 70° to the horizontal. A second pair of arrays was placed at right angles to the west-east arm and inclined at 70° as shown. The two acoustic beams intersected at 275 m above a point, 100 m from each pair of arrays. Alternate angle-of-arrival and Doppler measurements were made on the north and east arms to obtain measurements of both components of the wind. This was done by interchanging the north and east pairs of arrays before the transmitted pulse.

Angle-of-arrival measurements using the north arrays gave the west-east component of the wind, while Doppler measurements using these arrays gave the north-south component of the wind. Similar measurements using the east arrays gave the north-south and west-east components of the horizontal wind. The total velocity and direction of the wind were computed from these measurements and plotted with the wind velocity and direction at 10 m (fig. 8). The measurements were made during the night when the wind was strengthening prior to the passage of a cold front. There was a deep shear layer evident on the sounder record to about 300 m. The range gate was set within this shear at 120 to 150 m. The similarity between the three velocity traces is apparent. Note that the mean wind (V_θ) to 135 m (angle-of-arrival measurement) was closer to the surface value than it was to the velocity (V_r) at 135 m (Doppler measurement), suggesting that the wind shear was not constant up to 135 m but increased with height toward this level.

There was very good agreement between the three methods of measuring the wind direction.

Scatter diagrams of total wind velocity were constructed from 1.5-min. samples as shown in figure 9(a), (b), and (c). The correlation between each of the velocity measurements is shown to be very close.

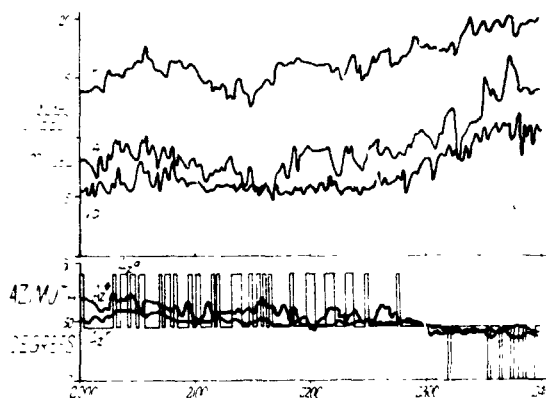


Figure 8.— Correlation between three methods of measuring total wind velocity and direction, 2000-2400 C.S.T. April 5, 1971.
(a) Wind velocity — V_r , V_θ , and V_a .
(b) Wind direction — A_z^r , A_z^θ , and A_z^a .

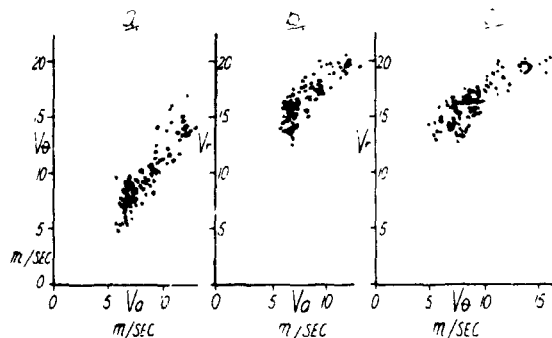


Figure 9.— Scatter plots of data from figure 8(a).
(a) V_θ and V_a ; (b) V_r and V_a ; (c) V_r and V_θ

VERTICAL GRADIENT OF WIND VELOCITY

The oblique monostatic configuration of the sounder (fig. 2(b)) was used to measure the vertical gradient of the wind velocity. Two arrays were mounted on the horizontal turntable and inclined at an angle of 20° to the horizontal (fig. 1). The turntable was rotated until the boresight direction of the arrays pointed directly downwind. The height of the Doppler range gate was switched alternately between two adjacent levels, 30 m apart, to obtain the wind velocity at each level. The difference between these two Doppler measurements of horizontal wind velocity was recorded to obtain the gradient over a thickness of 30 m.

The facsimile record shown in figure 10 was obtained when a wind shear layer developed within a radiation inversion during the night of April 5, 1971. The wind velocity at the surface increased from 4.5 m/sec at 2100 CST to 6.5 m/sec at 2200 CST and the temperature decreased by 0.5°C . During the same period, it will be seen that the top of the shear layer decreased from 250 to 150 m. Twenty-minute averages of the wind velocity at each level were obtained as shown in figure 10 before the mean height of each gate was decreased to keep within the shear layer. The average value of the wind velocity recorded in each range gate for each 20-min period was plotted above the sounder record in figure 10. The wind velocity recorded by the anemometer at 10 m above the ground was also plotted. The vertical gradient measured by the sounder was 2 m/sec/100 m for each 20-min sample.

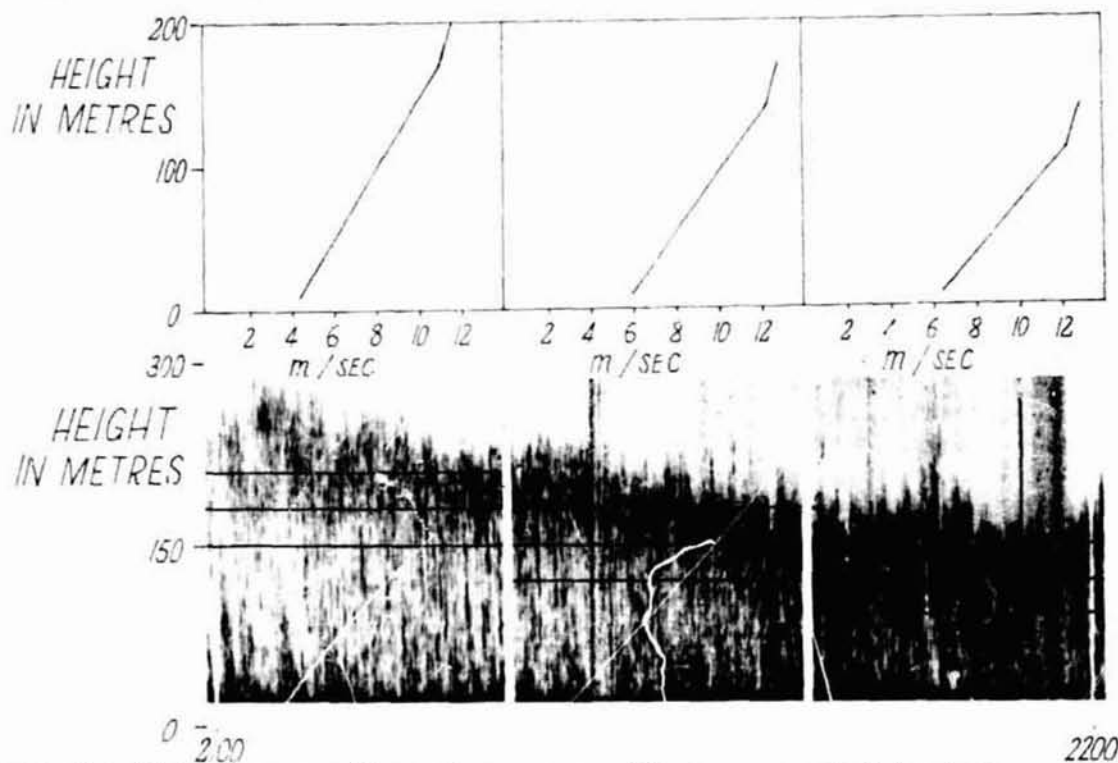


Figure 10.— *Top* — Wind profiles constructed from surface anemometer and Doppler measurements of wind velocity.
Bottom — Simultaneous inclined monostatic sounder record showing descending shear layer, 2100-2200 C.S.T. April 5, 1971.
 Height of range gate shown by parallel lines on record.

The wind velocity could not be measured at more than two range gate levels with the present equipment, but it is clear that if more levels were available, or if the meteorological conditions were sufficiently stationary, it would have been possible to measure the wind profile from the surface to the top of the wind shear layer.

CONCLUSIONS

The potential of the acoustic sounding technique for atmospheric research and operations has been appreciated for some time, but much of the detail of atmospheric motions contained in the sounder record still remains to be interpreted. A fundamental difficulty is the inability to make meteorological measurements in the same volume of space that is interrogated by the sounder, and progress to date has been limited for this reason.

Angle-of-arrival and Doppler techniques were applied to the measurement of wind velocity as a first step toward developing the acoustic sounder into an independent meteorological tool. The preliminary results of experiments were very encouraging. Good correlation was obtained between the wind velocity and direction measured by two independent, acoustic sounding techniques, and the wind velocity and direction recorded by an anemometer at the surface. The agreement between the three measurements of wind direction was particularly good when the facsimile record showed the atmospheric structure to be relatively simple.

The ability to measure the wind field in the first kilometer or so of the atmosphere, remotely from the ground, seems to be within reach. The feasibility of measuring the mean profile of temperature in the lower atmosphere by angle-of-arrival measurements in the vertical plane is being examined.

ACKNOWLEDGMENTS

The generous assistance given by J. R. Pollard, A. R. Mahoney, and M. V. Buggy during the preparation of this paper is gratefully acknowledged.

DISCUSSION

Peterson: How do you obtain profiles?

McAllister: Although we are range gating at one level, we could gate at different levels for profiling.

Reinisch: Could you sound at different frequencies to sound different heights?

McAllister: The scatter cross section is a slow function of frequency, $f^{1/3}$. Absorption of sound is a more sensitive function of frequency and relative humidity.

Parker: To do the inverse problem you must first be able to do the direct problem. Given a wind profile do you have the theory to predict the experimental data?

McAllister: Yes, but the turbulent medium is a complication, although a second-order effect.

Wright: Since you are depending on scatter, rather than reflection, selection of the turbulence scale sizes should be based on the wavelengths used. This may have a strong frequency dependence. Have you done multiple wavelength studies?

McAllister: We have looked as high as 1.5 KHz. Qualitatively there is no change from 800 Hz. Since the frequency dependence is small, we would need a factor of 10 change in frequency. Then absorption changes would be important.

Reinisch: If amplitude measurements at different frequencies were made, information on the type of scatterer would be revealed. Then perhaps one frequency could be used for profiling.

Bojarski: Do you have phase data?

McAllister: We have relative phase data between the receiving arrays.

McClatchey: In our problem — turbulent scatter of laser radiation — ways of describing the turbulence is a key unknown. You need detailed wind and temperature structure. Therefore, the turbulence should be defined statistically, and an inversion technique is needed that yields statistical parameters.

McAllister: Provided the process is stationary, which is often not true.

ATMOSPHERIC RADAR SOUNDING

R. K. Crane

Lincoln Laboratory
Massachusetts Institute of Technology
Lexington, Massachusetts 02173

N73-11593

ABSTRACT

Monostatic and bistatic radar techniques for the measurement of the structure of volume targets in the troposphere and lower stratosphere are reviewed. The targets considered are thin turbulent layers in the lower stratosphere and rain in the troposphere. Examples are given of measurements of both types of atmospheric scatterers taken with the Millstone Hill L-band radar and the Avon-to-Westford bistatic scatter systems.

The measurements of scattering from thin turbulent layers show that layers are generally detected at or near the tropopause, and in 31 out of 34 sets of measurements, layers were detected above the tropopause in the lower 10 km of the stratosphere. The layers all appeared to be widespread, covering hundreds of km in horizontal extent. The intensity of the layers varies in time and space with significant variations in tens of minutes and horizontal distances the order of 20 km. The threshold for turbulent layer detection corresponds to an equivalent C_n^2 thickness product of $10^{-13} \text{ m}^{1/3}$ at a range of 100 km and for layers with less than 10^3 m thickness.

The measurement of scattering by rain shows that in the New England area both convective and widespread rain consists of a number of small cells. On average, the cells appear to have a half-intensity width of 3 to 4 km as measured with a radar system with a 1.8 km resolution cell size for cells at 100 km range.

INTRODUCTION

Conventional radars have been used for many years for the study of precipitation processes [Atlas, 1964]. In the past decade, as a result of advances in radar techniques and devices, high-power and high-resolution radars have been made available to radar meteorologists for use in probing the atmosphere. These radars have been useful in the direct probing of two different atmospheric processes: turbulence as revealed by fluctuations in the index of refraction of the atmosphere and rain. High-power, high-resolution radars such as the Millstone Hill radar or the Wallops Island radars are required for the study of turbulence because of the relatively small radar cross section per unit volume produced by refractive index fluctuations. High-resolution radars are also useful for the study of rain since rain generally occurs in cells with limited horizontal extent.

The radar measurements of turbulence and rain presented here were made with the Millstone Hill L-band radar. The data are all the results of direct measurements in that the object of the measurement program was the cross section per unit volume as presented. Inversion techniques were not used except in their simplest form in the preparation of the data. The results of the measurements are of interest because they illustrate the spatial variability that is to be expected in meteorological processes. The rain measurements have additional significance for users of passive microwave systems since rain is a source of error in atmospheric sounding for the construction of temperature or humidity profiles.

THE RADAR SYSTEM

The parameters of the radar system are given in table 1. The radar system has a resolution volume at 100 km that may be approximated by a cylinder with a 1.8 km diameter and a 1.8 km length. The system has sufficient sensitivity to detect refractive index fluctuations due to turbulence in the lower stratosphere and fair weather cumulus clouds at 100 km.

TABLE 1.- PARAMETERS OF THE MILLSTONE HILL L-BAND RADAR

Frequency	1.295 GHz (23.2 cm wavelength)
Antenna	84-foot parabola with Cassegrain feed
Antenna gain	47.1 dBS
Beamwidth	0.6° between half-power points
Polarization	Right-hand circular transmitted Left-hand circular received
Transmitted power	4 Mw peak (continuously monitored)
Pulse length	10μ sec
Pulse repetition rate	20 per second
Receiver bandwidth	80.5 KHz (12.4μ sec matched predetection filter)
Data processing	Analog to digital conversion of IF sine and cosine channels every 10μ sec
Detection	Square Law by computer operation
System noise temperature	290°K (includes atmospheric and ground effects averaged over 0-30° elevation angle)
Overall system feed and line losses	1.7 dB
Matched filter processing loss	1.4 dB
Single pulse C_N^2 value for unity signal-to-noise ratio	$2 \times 10^{-16} \text{ Km}^{-2/3}$ at 100 km
Minimum detectable layer C_N^2 value with horizontal averaging and average-noise subtraction	$1 \times 10^{-16} \text{ Km}^{-2/3}$ at 100 km
Single pulse Z value for unity signal-to-noise ratio	$1 \times 10^{-3} \text{ mm}^6/\text{m}^3$ at 100 km

Rain and refractive index fluctuations are extended targets in space. However, these targets are inhomogeneous, and the radar therefore provides an estimate of the cross section of the target as averaged over the radar beamwidth and one-half the pulse length. The radar data presented here have been scaled by the effective integration volume of the radar system and are reported as cross section per unit volume.

The radar measurements reported have been made as a part of a program of microwave propagation research, which also includes measurements of bistatic scatter by rain and refractive index fluctuations and passive radiometric measurements of rain [Crane, 1970 a; Crane, 1968]. The bistatic measurements were made at X-band using a 145 km scatter path between a transmitter in Avon, Connecticut, and the Westford Communications Terminal; the parameters of this system are given in table 2. The radiometric measurements were made using a X-band radiometer and the Haystack antenna system; the parameters of this system are given in table 3. Measurements with the radar and at least one of the other two systems were always made simultaneously with the receiver antennas scanned in synchronism.

MEASUREMENTS OF REFRACTIVE INDEX FLUCTUATIONS

The models for scattering by refractive index fluctuations presented by Tatarski [1961] indicated that bistatic scattering by turbulence should be one of the dominant mechanisms for transhorizon scatter propagation. The same model also predicted that high-power radar systems operating in the 10- to 50-cm wavelength region should detect scattering from turbulence. The high-power S-band radar system at Wallops Island provided some of the first data to verify the prediction [Hardy et al., 1966]. Measurements with the Millstone Hill L-band radar [Crane, 1970 b] and other installations have also verified the existence of scattering from turbulence.

The model presented by Tatarski related the scattering cross section per unit volume to C_N^2 , the structure constant for turbulent fluctuations of refractive index and a measure of their intensity:

$$\beta_1(\phi) = 0.378\lambda^{-1/3} C_N^2 \left(\sin \frac{\phi}{2} \right)^{-11/3}$$

$$\beta_2(\phi) = \beta_1(\phi) \cos^2 \phi \quad (1)$$

where

- β cross section per unit volume
 \perp, \parallel polarization perpendicular \perp and parallel \parallel to the plane of scattering
 λ wavelength
 ϕ scattering angle

The model shows that for forward scatter, ϕ small, the scattering cross section per unit volume is much larger than for radar measurements, $\phi = \pi$. This result is shown in the experimental data.

Tatarski went further and proposed a meteorological model for the estimation of C_N^2 given profiles of wind, temperature, and water vapor. In its simplest form, the model indicates that C_N^2 is, to first order, proportional to the square of the gradient of absolute humidity in the lower atmosphere or of potential temperature at heights at or above the tropopause and inversely proportional to the two-thirds power of the magnitude of the vector wind shear. Computations based on this model and the use of measured profile data show that in the region at and above the tropopause, the C_N^2 values generally are higher in regions with strong gradients in potential temperature. The radar measurements presented in figures 1 through 7 are consistent with this observation. The apparent thinness of the thin turbulent layers as shown in these figures indicate that the regions with strong potential temperature gradients are rather thin and profiles of temperature will have marked small-scale changes in their average properties.

The radar measurements shown in figures 1 through 7 were made with the system described in table 1. This system differs from that typically used for meteorological measurements due to the use of a general-purpose digital computer as an integral part of the receiver. The computer detects the radar returns, incoherently averages the returns, and calibrates the receiver system using signals from a noise tube that is fired after each data-taking operation, once per transmitted pulse. To improve the target-detection capability of the radar, the recorded data consists of the incoherent average of the signal-plus-noise minus the average of the noise. The recorded data for each elevation scan is further computer analyzed to provide incoherent averages of the received signal over 22.5 km horizontal distance interval at a constant height. The results of a series of the horizontal averages at different heights is then assembled into a profile of C_N^2 versus height as shown in figure 1; C_N^2 as used here is found by solving Eq. (1) given the radar measurement of β . The corresponding computer-generated RHI display is shown in figure 2. As an aid in data editing, selected elevation scans are also recorded in the conventional manner by photographing a PPI display. The PPI display that corresponds to the scan used to generate figures 1 and 2 is shown in figure 3.

TABLE 2.—AVON-WESTFORD X-BAND SCATTER PATH PARAMETERS

Frequency	7.74 GHz (3.86 cm waves)
Antenna 1	60-foot parabola with Cassegrain feed
Aperture efficiency antenna 1, %	40 percent
Beamwidth antenna 1	0.15° between half-power points
C 2	1.48
Polarisation antenna 1	Left-hand circular received
Antenna 2	6-foot parabola with prime focus feed or standard gain horn
Gain antenna 2	37.5 dB for 6 foot 18.2 dB for horn
Beamwidth antenna 2	1.5° for 6 foot 23° for horn
Polarisation antenna 2	Vertical or horizontal linear transmitted
Transmitter power	Variable to 500 w
Transmitted signal	cw with frequency stable to 1 part in 10^{10} per day
Receiver	Phase lock
Receiver bandwidth	500 Hz
Receiver noise temperature	250°K (includes atmosphere and ground effects)
Maximum detectable transmission loss	220 dB with 6 foot 200 dB with horn
Path length	145 km
Data processing	Received signal AGC voltage and local oscillator frequency sampled 20 times per second

TABLE 3.—RADIOMETER PARAMETERS

Radiometer Parameters	
Frequency	7.74 GHz (3.86 cm).
Antenna	120-foot parabola with Cassegrain feed.
Polarisation	Left-hand circular.
Beamwidth	0.15° between half-power points.
Receiver Bandwidth	16 MHz pre-detection.
Detector	Switched comparison-type radiometer.
Integration Period	0.3 seconds.
System Noise Temperature	60°K.
Temperature Measurement Uncertainty	0.03°K.

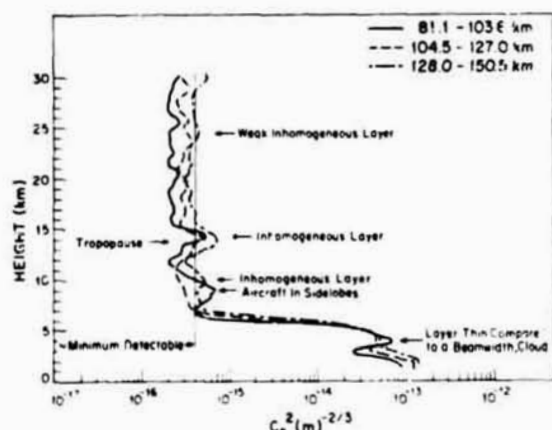


Figure 1.— C_h^2 profiles for 2 August 1968, 1720 GMT, 233° azimuth made using the Millstone Hill radar.

layer at 14 km in figure 1 were 100 m thick, the actual C_h^2 value for the 81.1 to 103.4 km average would be $4 \times 10^{-15} \text{ m}^{-2/3} (2 \times 10^{-16} \text{ cm}^{-2/3})$. A homogeneous layer that is thinner than 1 km would have different C_h^2 values reported because the percentage of the beam filled depends on the distance from the radar. For thin layers the ratio of the measured layer intensities are 1:1, 25:1, 50:1, respectively, for the 81.1- to 103.4-, 104.5- to 127.00-, and 128.0- to 150.5-km distance intervals. The relative magnitudes of the C_h^2 profiles for the layer at 4 km show that if the layer were homogeneous, it would be thinner than 1 km and the C_h^2 value reported is a lower bound on the layer strength.

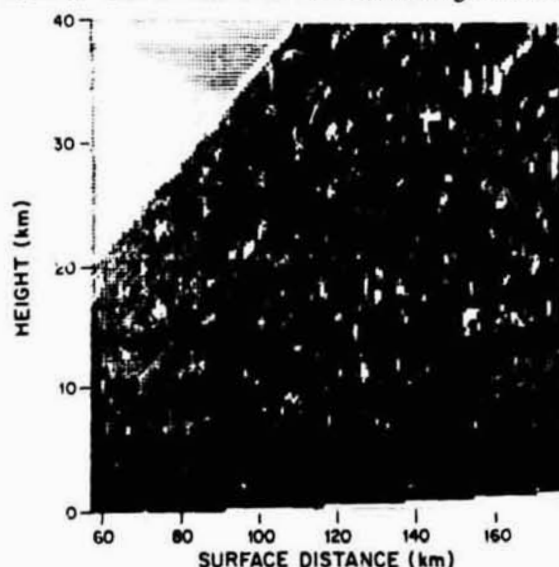


Figure 2.— Computer generated RHI display for 2 August 1968, 1720 GMT, 233° azimuth.

80- to 150-km measurement range was dictated by ground clutter problems. Fixed targets and aircraft at ranges between 80 and 150 km must be manually detected using the computer-generated RHI displays. Data contaminated either by interference or extraneous targets were rejected.

The L-band radar used can readily detect turbulent layers with a C_h^2 height product greater than $1 \times 10^{-16} \text{ m}^{-2/3} \text{ km} (5 \times 10^{-18} \text{ cm}^{-2/3} \text{ km})$. Unfortunately, a cloud of the same dimension and a Z of $6 \times 10 \times 10^{-4} \text{ mm}^6/\text{m}^3$ (see Eq. (2)) for liquid particles or a Z of $3 \times 10^{-3} \text{ mm}^6/\text{m}^3$ for ice particles will give the same return. This corresponds to a cloud of liquid particles with a liquid water content the order of 0.005 g/m^3 . The detection of thin turbulent layers

The radar system was operated to record data at distances between 80 and 150 km from the radar and heights between 5 and 30 km. At 100 km range and at the elevation angles used, the effective integration volume of the radar is approximately 1.1 km in height, 1.1 km in horizontal distance normal to the plane of the radar and scattering volume, and 1.1 km in horizontal distance in the plane of the radar and scattering volume. The height resolution capability of the radar at 100 km is 1.8 km, which corresponds to the 10 dB beamwidth of the antenna. This indicates that thin layers closer together than 1.8 km in height will not be resolved and, for layers thinner than 1.1 km, the measured layer strength will be a function both of the layer thickness and C_h^2 . Since most of the layers probed by a U-2 aircraft used in one part of the measurement program were the order of 100 m thick, it is best to view the layer strength measurements as providing the value of the C_h^2 height product. If the

The radar system has been calibrated so that the measured C_h^2 value for a layer that fills the beam is within $\pm 1.5 \text{ dB}$ of the actual value. The stability of the radar, the characteristics of the radar target, and the pulse-by-pulse calibration system used in making the measurements provides a relative rms error of less than $\pm 0.7 \text{ dB}$. The major sources of error in making the C_h^2 profile measurements are interference from other radars operating in the same frequency band, reflections from fixed ground targets detected through the side lobes of the antenna pattern, and reflections from aircraft detected through the side lobes. Each error source is monitored during the measurement program. Interference is monitored by the relative magnitude of the variance of the receiver noise temperature which is continuously measured. The

is therefore complicated by masking by cloud layers. For this reason, only layers in the clear atmosphere may be detected. Since a single-frequency radar is used, auxiliary means must be used to determine if clouds are present. For layers near the tropopause, the presence of cirrus clouds must be determined. The separation between thin turbulent layers and cirrus cloud layers has been made using several criteria. If the layer is detected above the tropopause in a region where the radiosonde data show a definite lack of moisture, the layer is assumed to be turbulent. If the layer is thin, on the order of the resolution volume of the radar, and is thin throughout the duration of the test, it is assumed to be turbulent. If no cirrus clouds are visually observed by pilots, ground-based observers, and the available TPQ-11 cloud radars at the height of the layer, the layer is assumed to be turbulent. Whenever doubt existed as to the origin of the radar return, the layer was assumed to be caused by clouds.

Using the above layer-detection criteria, the data from 34 days of measurements randomly distributed between January 1968 and August 1968 were examined for the existence of thin turbulent layers at heights above the tropopause. The height of the tropopause was determined using the radiosonde data closest both in time and space to the measurement volume. Two types of layers are reported, persistent and transient. A persistent layer is a layer at a specific height that is present in more than half the elevation scans. The measurements were made by making two elevation scans at a particular azimuth; then moving the antenna to a new azimuth for another set of elevation scans. The azimuth used for the measurements changed from day to day. Typically a series of measurements were made at five separate azimuths spaced over a 180° interval interlaced with a series of measurements at azimuths spaced over a 20° azimuth interval. With the azimuth stepping method used for data taking, a layer had to be present over a range of azimuth values to be classified as persistent.

A series of measurements made on May 28, 1968 is shown in figures 4 and 5. In this set of data, a persistent layer existed at the tropopause level. The layer at 14-km height is also classified as persistent since it is present in more than half the elevation scans. As shown by the azimuths for which the scans were made, the persistent layers are not confined to a single azimuth sector but are detected at a wide variety of azimuths. These layers extend more than 300 km



Figure 3.— RHI display for 2 August 1968, 1720 GMT, 233° azimuth made using the Millstone Hill radar.

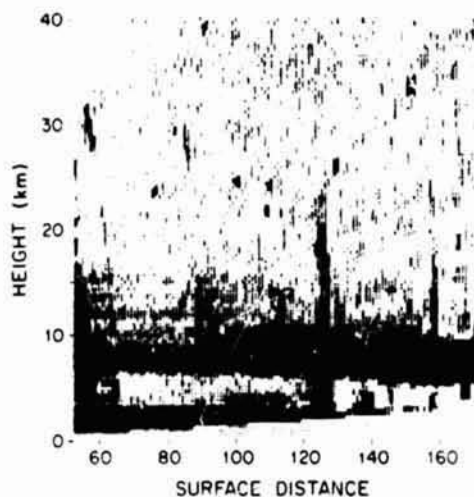


Figure 4.— Computer generated RHI display for 28 May 1968, 1439 GMT, 120° azimuth.

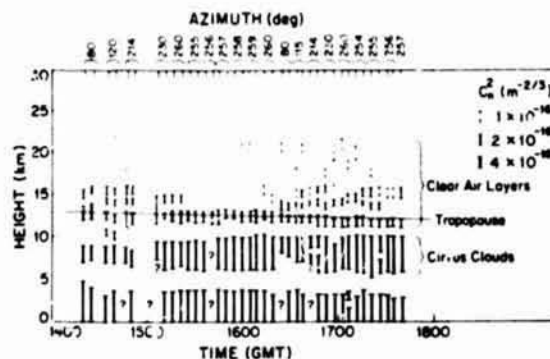


Figure 5.— Radar detected atmospheric layer structure for 28 May 1968.

horizontally. The intensity of the layers, however, changes in time and position. The intensity of the layer at tropopause height in figure 5 fluctuates in time with approximately a 1-hr period. The data on figure 5 also show three transient layers at heights above the tropopause. Again, these transient layers tend to be widespread in horizontal extent but weaker in intensity.

The data for each of the measurement days are presented in table 4. The results are given by day. The table also gives the number of layers per 2.5-km height interval for heights above 10 km and the total number of layers detected above the tropopause for each day. The data generally show a layer just above the tropopause. The layer probably is associated with the tropopause but, due to the uncertainty in both the existence and location of a tropopause, it is hard to make a definite judgment about such an association. The feature of the radiosonde profiles most often associated with the layers that occur above the tropopause is a temperature inversion.

TABLE 4.- TURBULENT PERSISTENT (P) AND TRANSIENT (T) LAYER DETECTIONS

Table IV Turbulent Layer Detections													
Date	Time GMT	Max. cloud level km	Tropopause Height km	Layers above tropopause		Layers from RHI Data							
						10 - 12.5 km		12.5 - 15 km		15 - 17.5 km		17.5 - 20 km	
				P	T	P	T	P	T	P	T	P	T
1/9	1800-2100	6	12.9				1						
1/11	1420-2050	-	11.0		1		1		1				
2/5	1720-2130	2	6.5		1		1						
2/9	1400-2050	6	10.0		1					1			
2/13	1340-1850	6	7.3	1	4	1	1		1			1	1
2/19	1810-2100	5	7.6		2		1					1	
3/7	1410-2050	1	9.5		6		1		1			1	2
3/15	1410-1840	6	11.5	1	2		1		1				
3/20	1400-2100	10	11.8		2	1			1				
3/21	1730-2100	10	12.6		1				1				
3/22	1340-1700	10	12.5	1	2				1			1	
4/4	1800-2130	12	12.5		1				1				
4/15	1330-2120	9	11.5		2	1			1				
4/24	1600-1800	9	12.0		1	1			1				
5/6	1310-1900	2	12.7						1				
5/16	1310-1710	9	12.5	1		1			1				
5/20	1440-1830	2	10.5	1	1	1			1				
5/28	1400-1950	9	12.5	1	5	1			1			1	
5/29	1300-2030	11	12.3		5		1		1		2	1	1
6/3	1410-1950	10	10.4		2	1	1		1				
6/5	1550-1930	9	12.8		1	1			1				
6/12	1350-1720	12	14.8		2				1		2		
6/27	1230-1630	11	14.8		3		1					1	1
6/28	1240-1600	8	12.7	1					1				
7/29	1430-2020	3	13.8	1	3		1	1				1	1
7/30	1340-1950	11	12.4		2				1			1	
7/31	1230-2020	11	12.8	1	1		1	1				1	
8/1	1330-0100	9	14.5		2		1	1			1	1	
8/2	1400-2020	5	13.8		1			1				1	
8/5	1220-2030	8	13.1		1				1				1
8/6	1240-2100	9	14.2		2	1			1			1	1
8/7	1230-0100	9	14.4						1				
8/8	1240-2020	7	14.1		2	1	1	1			1		
8/9	1300-2020	9	12.6			1	1	1			1		
Total 34	Total of 107 layers detected above 10 km			9	57	12	14	18	13	3	20	14	2
				66	28*	31	23	14	7	4			

* Total is low due to rejection of any layer with clouds.

Transient layers have been detected at heights up to 10 km above the tropopause. The maximum height for a layer detection is 25 km. Weaker layers probably exist above this height but the radar system sensitivity is not high enough to detect them. The transient layers may also appear persistent to a more sensitive radar.

Simultaneous bistatic scatter measurements were also made of scattering from refractive index fluctuations using the Avon-to-Westford scatter system outlined in table 2. The radar measurements and bistatic scatter measurements are compared in figures 6 and 7. Figure 6 shows the C_n^2 profile as measured by the radar. Figure 7 shows the results of scanning the Westford antenna in elevation while pointing along the great circle path. The measurements shown by the points show the layered structure. The arrows point to the locations of layers that also exist in the radar data. The horizontal dotted lines show the transmission loss value, or ratio of transmitted-to-received power, (P_t/P_r) , computed using the bistatic radar equation, the scattering angle dependence given in Eq. (1), and the radar estimate of C_n^2 . At elevation angles below 5° agreement is evident between the computed and measured values. Above 5° , coupling via the sidelobes of the Westford antenna becomes important and the computation model breaks down. The data show, however, that the model proposed by Tatarski that relates C_n^2 and β is useful in the interpretation of scattering from thin turbulent layers.

MEASUREMENTS OF RAIN SHOWER STRUCTURE

The scattering cross section for rain is related to Z , the sum of the sixth power of the drop diameters by Rayleigh scatter theory:

$$\beta_{\perp}(\phi) = \frac{\pi^5}{4} |\kappa|^2 Z$$

$$\beta_{\parallel}(\phi) = \beta_{\perp}(\phi) \cos^2 \phi \quad (2)$$

where $|\kappa|^2 = |(e-1)/(e+2)|^2$, e = dielectric constant of water, and the other parameters are defined in Eq. (1). Equation (2) is valid for frequencies below 10 GHz. The parameter Z is a function only of the drop-size distribution and may be considered to be a meteorological parameter that is descriptive of rainfall intensity. The relationship between Z and the rainfall rate R is a statistical one that depends on the properties of drop-size distributions. The relationship between Z and R normally used by radar meteorologists is $Z = 200R^{1.6}$ when Z is expressed in mm^6/m^3 and R in mm/h .

Millstone Hill L-Band Radar
 C_n^2 Profile Measurement

2 Aug 1968 1630-1636 GMT
220° Azimuth

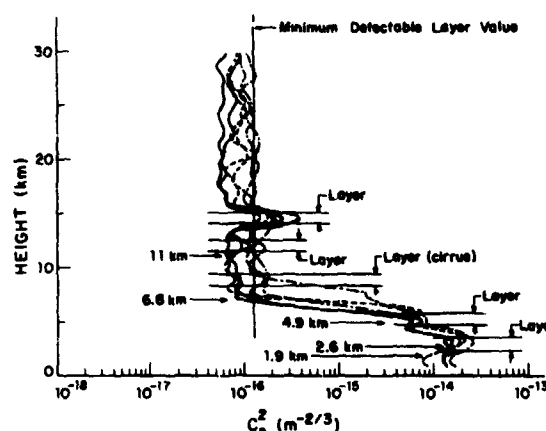


Figure 6.— C_n^2 profiles for 2 August 1968, 1630 GMT, 220° azimuth made using the Millstone Hill radar.

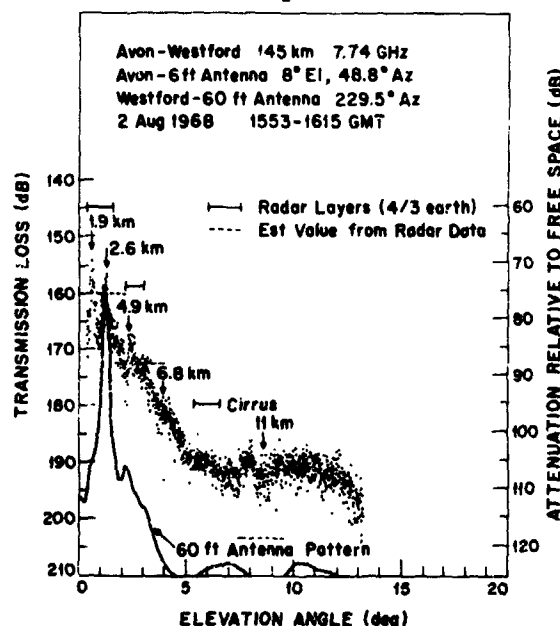


Figure 7.— Transmission loss vs receiver elevation angle, 8.0° transmitter elevation angle.

Radar measurements of rain generally show a marked horizontal inhomogeneity. Although propagation specialists and climatologists often consider two rain types, stratiform and convective (or showery), measurements with high-resolution radars generally show that both rain types are inhomogeneous and quite showery.

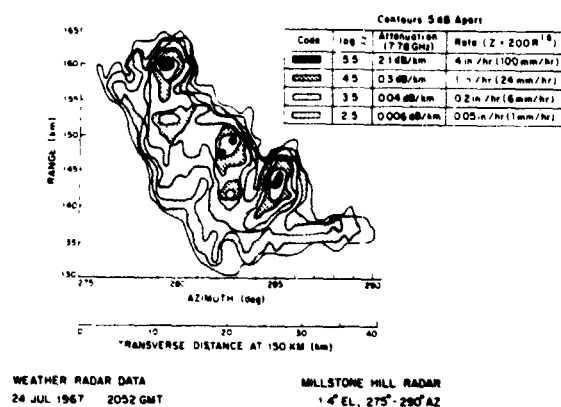


Figure 8.— Weather radar map for New England showers.

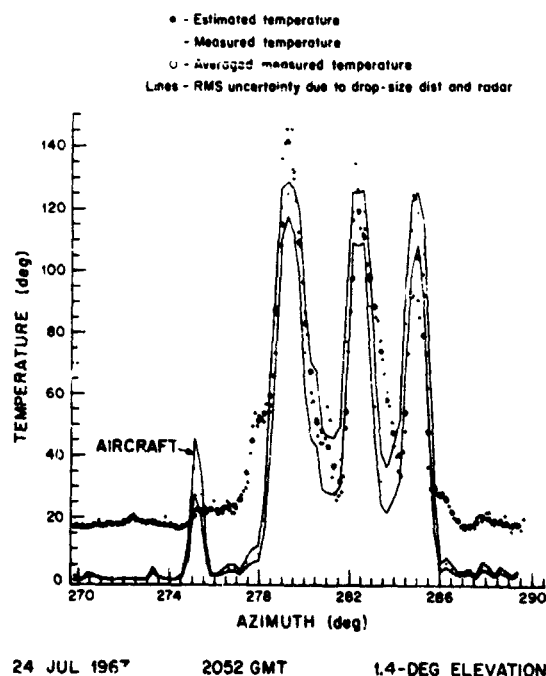


Figure 9.— Estimated and measured antenna temperature relative to a 2° elevation angle clear sky value.

large number of comparisons were made with no intervening obstacles. The results of the latter set of measurements show agreement with the Rayleigh model within the measurement accuracies of the radar and scatter systems.

The showery structure of rain is also evident in radiometric measurements of rainfall made using the Haystack radiometer system described in table 3. Radiometric observations were made by scanning the antenna across rain showers of interest. The radiometric measurements made simultaneously with the radar measurements presented in figure 8 are presented in figure 9. The points labeled as measured and as averaged measured temperature are measured changes in antenna temperature from the clear sky value for an elevation angle of 2.0°. The solid lines enclose the ± 1 standard deviation estimate of antenna temperature change based on computations using an approximate solution to the radiative transfer equation and estimates of attenuation cross section per unit volume based on the radar measurements. The relationship between attenuation cross section and scattering cross section was derived statistically from a large number of computations using measured drop-size distributions [Crane, 1968]. The data presented in figure 9 are the direct measurements of antenna temperature changes — not changes in sky or brightness temperature. The true value of sky temperature change are higher and may be determined by dividing the listed temperature values by the integrated antenna efficiency, which is 0.6 for the elevation angles used. The measurements show reasonable agreement between measurements and estimates based on radar data.

Bistatic measurements of rain shower structure were made using the scatter system described above. The objective of the measurement program was to measure the scattering angle dependence of natural rain. The measurements also depict the showery nature of rain. Figure 10 shows a radar measurement similar to that in figure 8. The simultaneously made bistatic scatter observations are shown in figure 11. Note the rapid spatial change of rain intensity. The cells evident in figure 10 are marked by arrows in figure 11. The horizontal lines give the estimated transmission loss based on the several scattering models listed and the use of the bistatic radar equation. The only agreement shown here is at an azimuth angle of 215°. At the other azimuth angles, obstacles existed between the transmitting antenna and the scatter volume and agreement was not obtained. In a recent set of measurements [Crane, 1971], a

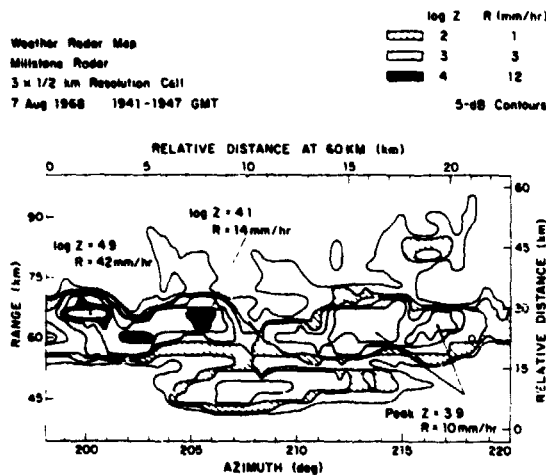


Figure 10.- Weather radar map.

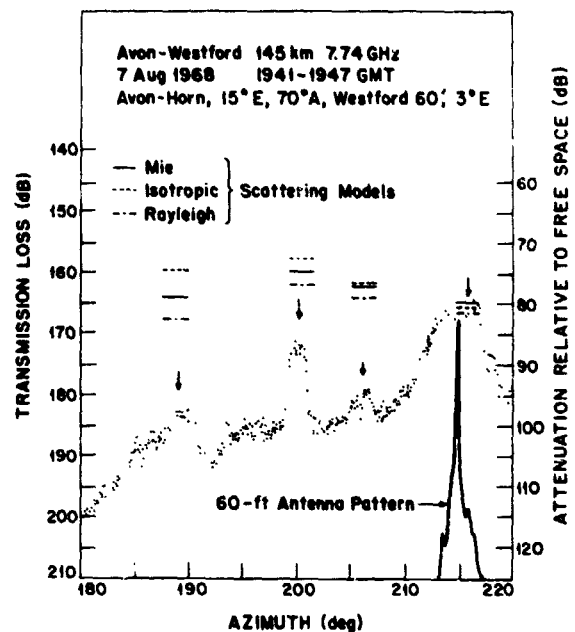


Figure 11.- Transmission loss measurements vs receive antenna azimuth, 3.0° receive elevation angle.

SUMMARY

A series of measurements of the structure of two atmospheric parameters detectable by radar techniques were presented. The measurements were obtained as part of a program of microwave propagation research which has as its object the testing of models used to predict propagation phenomena. The measurements show the inhomogeneity typical of atmospheric parameters. Experience obtained in the course of the research program has indicated that as radars are developed that have higher resolution volumes, measurements of atmospheric phenomena will show significant changes in the intensity of the phenomena over the smallest distances resolvable.

ACKNOWLEDGMENT

This work was sponsored by the National Aeronautics and Space Administration under Contract NAS5-21544.

DISCUSSION

Chang: Are you familiar with Tatarski's 1967 book in which he shows C_n is a function of temperature and wind velocity and also extends the equation over a wide range?

Cran: I have not seen the translation but there is not a big change from his earlier work. Some people, in fact, doubt his extension. Other recent Russian work is even more important but difficult to apply. They show that details within a volume tend to govern C_n and we can't get this from profile measurements. All we can find is the region to expect a layer but not its strength.

Unidentified speaker: How did you obtain estimates of rain drop-size distributions and how critical are they?

Cran: From a wide number of sources, predominantly camera surveys of the Illinois State Water Survey. New AFCRL measurements improve on these. We can classify rain by type and get better estimates of attenuation given reflectivity and rain rate. Our accuracy is about 30% standard error of estimate in attenuation cross section per unit volume due to drop-size distribution affects. The AFCRL results may reduce this number.

FINE STRUCTURE OF THE LOWER ATMOSPHERE

AS SEEN BY HIGH RESOLUTION RADAR

J. H. Richter, E. E. Gossard, D. R. Jensen

Propagation Technology Division
Naval Electronics Laboratory Center
San Diego, California 92152

1 N73-11594

ABSTRACT

A ground-based vertically pointing FM-CW radar (frequency modulated, continuous wave) is described that permits remote probing of the refractive index structure in the troposphere. The radar has the characteristics of extremely high sensitivity, ultrahigh range resolution, and close minimum detection range without clutter. The sounder routinely detects layer structures in the lower troposphere. These layers are always associated with gradients in the vertical refractive index profile, and are frequently very thin, approaching the resolution of the radar (1 m). Very often they are perturbed by wave motions. Examples of various wave patterns are presented, and an explanation is given for organized substructures frequently superimposed on larger scale wave motions.

INTRODUCTION

One important advantage of remotely probing the atmosphere is the capability for continuously monitoring the layer structure in depth. Pulsed radar has long been used in several countries to detect the height of such layers, but the lack of resolution has precluded the possibility of examining structural detail. Only thick diffuse layers were normally revealed, although these occasionally showed interesting wave-like structures. Two kinds of returns were observed: diffuse echoes from regions of large gradient of refractive index, and "dot angels" from insects and birds. Since the refractive index depends strongly on moisture and temperature, radar was evidently an important potential instrument for remotely probing the meteorological structure of the clear troposphere, if an improved radar system could be developed. Such a radar has been designed and built at NELC. It is a unique FM-CW radar (frequency modulated, continuous wave) that differs from previous radar sounders mainly in two features: extremely high range resolution (1 m) and low ground clutter. Its narrow beams are directed vertically, which permits us to observe features as low as 50 m from the ground. The sounder has been in operation since December 1968, and observations of great variety have been obtained. The radar observations are accompanied by both standard and special meteorological observations at the ground and by radiosonde soundings. The radar sounding system and some early observations have been described by Richter [1969].

DESCRIPTION OF RADAR

For the study of the fine-scale refractive index structure in the troposphere we need a radar sounder that combines the properties of high range resolution, high sensitivity, and close minimum range without clutter. The requirement of high range resolution capability would suggest the use of pulse compression techniques. However, this approach would make it difficult to meet the additional requirement of close minimum range, for no signal can be received during the transmission of the uncompressed pulse (and usually several pulsewidths after). The reason that the minimum range is such an important factor for the sounder is that low level inversions frequently produce scatter layers in the lowest 300 m of the atmosphere, and these layers can be studied most effectively by simultaneous meteorological observations by tower and tethered balloon instruments.

An FM-CW radar does not have the minimum range restriction and allows high range resolution. It has the additional advantage of a peak-power-to-average-power ratio of unity, which is very desirable from a technological viewpoint. It also permits a change of resolution without affecting the transmitted energy (which is not the case for pulse radars).

The frequency of the radar was chosen to be around 3 GHz as a compromise between higher and lower frequencies, each of which would have been desirable for different reasons. Thus, to achieve a narrow antenna beam with a small antenna diameter, it would have been better to select a higher frequency. However, we must also consider the reflectivity-wavelength dependence of the scattering volume. For specular surfaces, the power reflection coefficient increases as the thickness in wavelengths of the transition zone decreases, and we should thus want to use the largest possible wavelength. On the other hand, the preponderance of available evidence suggests that clear-air-scatter layers comprise inhomogeneities in refractivity, producing a reflectivity (η)-wavelength dependence of about $\lambda^{-1/3}$ [Atlas and Hardy, 1966; Hardy et al., 1966; Hardy and Katz, 1969]. Although this wavelength dependence is small, it suggests using the shortest wavelength, provided that it does not approach the limiting microscale (ℓ_m) of the refractivity perturbations. Atlas and Hardy [1966], therefore, suggest an optimum wavelength of about five times the limiting microscale. In weakly turbulent conditions, ℓ_m is of the order of 1 to 2 cm, and so a λ of 5 to 10 cm seemed optimum.

By means of theoretical calculations and experimental data [Saxton et al., 1964], we decided that with an equivalent pulse length, $h = 1$ m, and antenna gain of 35 dB, a ratio of 190 dB between average transmitted and received power should be adequate to detect most scatter layers in the lowest 2 km.

The linear frequency modulation is generated with a YIG tuned transistor oscillator, amplified to a maximum power of 150 W and radiated by the transmitting antenna. The signals are displayed on cathode ray tubes in an A-scope presentation, or an intensity modulation, or in a combination of both amplitude and intensity modulation (or A-I scope). The data are recorded by 35-mm, shutterless movie cameras with film speeds ranging from 1 cm/hr to 0.5 cm/sec.

The antennas consist of two parabolic dishes, 3 m in diameter, with waveguide feeds (fig. 1). For good isolation, the antennas were located in pits. At least 60 dB isolation is necessary to avoid saturation of the preamplifier. The

insides of the pits are lined with microwave absorbers to suppress reflections. The absorbing screens are also required to suppress echoes from nearby ground targets. The antennas are steerable within $\pm 2.5^\circ$ to optimize the common volume for a given height.

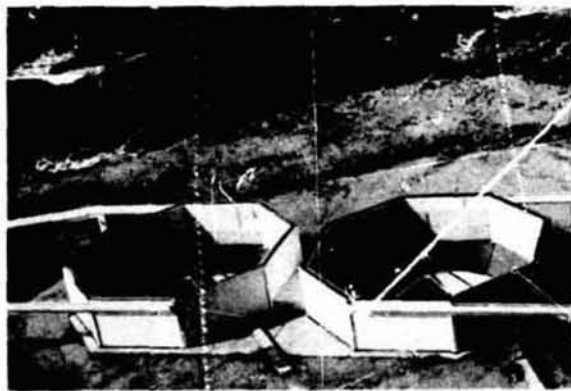


Figure 1.— Sounder antennas.

The performance characteristics of the radar are compiled in table 1. The basic components of the radar system allow operation in the entire 2- to 4-GHz range. The restriction to the narrower range, around 2.9 GHz, is determined by the standing wave ratio of the antenna feed and the balanced mixer. The variable range resolution is an important feature of the radar. In the case of volume scattering, signal amplitude changes continuously with resolution, thereby permitting significant conclusions to be drawn. The maximum range resolution of about 1 m was determined experimentally. Two targets 93 cm apart (simulated by delay lines) were separable by choosing $F = 219$ MHz frequency excursion.

The values for computing minimum detectable cross section and minimum detectable reflectivity are based on a transmitted power of 100 W and on a minimum detectable signal of -150 dBm. This minimum detectable signal was determined by using a delay line with a delay time equivalent to that of a target at a height of 167 m.

TABLE 1.— PERFORMANCE CHARACTERISTICS OF THE NELC RADAR SOUNDER

PARAMETER	VALUE	REMARKS
Power	150 W max	Typical power: 100 W
Center frequency	2.8 to 3.1 GHz	
Frequency excursion	Variable	Linear modulation
Range resolution	Variable	Maximum resolution: 1 m for 200-MHz frequency excursion
Sweep duration	50 msec	10 sweeps/sec
Receiver noise figure	5 dB	
Minimum detectable signal	-150 dBm	
Antenna gain	35 dB	
Antenna beamwidth	2.5 deg	
Isolation between antennas	105 dB	
Minimum detectable cross section at 1 km	$\cong 3.7 \times 10^{-6} \text{ cm}^2$	
Minimum detectable reflectivity at 1 km	$\cong 4.2 \times 10^{-15} \text{ cm}^{-1}$	For 1-m range resolution

The sensitivity of the radar for a point target is given by

$$\sigma_{\min} = 3.7 \times 10^{-6} r^4$$

where σ_{\min} is the detectable cross section in cm^2 , and r is the distance of the target in kilometers. Typical insects with cross sections of 10^{-3} cm^2 [Hardy and Katz, 1969], therefore, can cause dot echos 24 dB above noise at 1-km height. For distributed targets the minimum detectable reflectivity η_{\min} in cm^{-1} is

$$\eta_{\min} = 4.2 \times 10^{-15} r^2 / h$$

where r is the distance in kilometers and h is range resolution in meters. Typical values of η for a 10-cm wavelength are $10^{-16} - 10^{-14} \text{ cm}^{-1}$ [Hardy et al., 1966; Atlas and Hardy, 1966; Hardy and Katz, 1969]. It can be seen that even for the highest range resolution of 1 m, returns can be expected from as high as 1 km. To detect the weaker scattering regions at greater ranges, we should have to increase the effective pulse depth or resolution, provided of course that the scatter regions are thicker than the effective pulse.

Even though there is no basic minimum-range restriction inherent in this kind of radar, some practical limitations (which could be overcome, if necessary) reduce the sensitivity of the radar for targets closer than 50 m to the antennas. Reflections from an adjacent building give permanent returns for the first 50 m in height; therefore, targets must be capable of providing stronger returns or must change their position in order to be discernible. Also, the separation of the two antennas imposes a minimum height range for beam intersection. The antennas are separated by 5 m, and their beams intersect within 50 m above the ground.

RESULTS

Figure 2 is an example of layers that are observed by the radar system. The 50-min section of an intensity modulated trace shows a complex vertical structure consisting of thin, undulating scatter layers.

A sounding balloon carrying two radiosonde transmitters, which were modified to give continuous fast response temperature and humidity profiles, was launched at the radar site. The balloon was tracked from the ground to obtain wind information. The vertical profiles of temperature, relative humidity, wind speed, and wind direction are plotted in figure 2. The height scale for the balloon sounding and the radar return is the same. The refractive index $N = (n-1)10^6$ and the potential temperature are computed from the direct measurements.

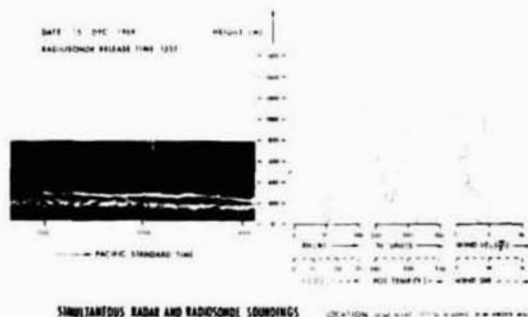


Figure 2.— Scatter layers observed by the radar and profiles of meteorological data obtained from balloon soundings. RH , relative humidity; $N = (n-1)10^6$, refractive index; t , temperature.

Figure 3 shows the approximate height and thickness of layer structure observed continuously from 6-8 January 1969. The agreement of backscatter zones and refractive index gradients is remarkable. The significance of this layer structure on radar coverage is illustrated in figure 5, in which a transmitter at a height of 100 ft. is assumed. The earth is drawn flat, which gives a straight ray an upward bend. From the ray trace picture we should expect a hole in coverage at a range of around 25 nautical miles. We also see from the ray trace picture that only the layer belonging to the steepest gradient influences the coverage significantly. This layer changed about 80 m in height between 1220 and 1240 PST, which could influence an accurate coverage prediction. This example emphasizes the need for precise and continuous knowledge of the refractive index structure, if accurate coverage predictions are required. Many more examples of layer structures with radiosonde sounding data are presented by Richter and Gossard [1970].

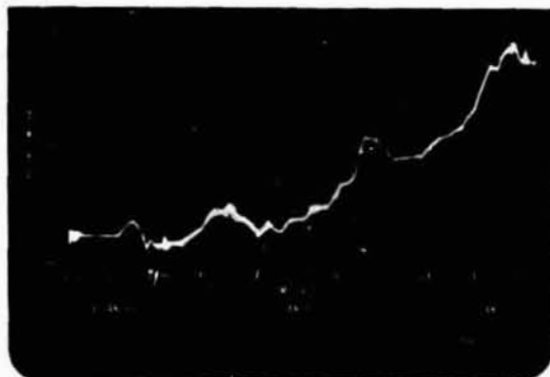


Figure 3.— Approximate height and thickness of layer structure observed continuously from 6-8 January 1969.

The temporal continuity of the radar observation technique may be used for presentations of the kind shown in figure 3, which is handdrawn to achieve the time compression necessary for illustration purposes. It shows height and thickness of the layer for a 2-day time period. No simple diurnal behavior of layer structure has been established so far, even though continuous observations for as long as 4 weeks have been studied. This may be due to superposition of other meteorological factors on a diurnal effect. It is also difficult in the case of multiple-layer structures to objectively identify and connect strata with strong fading characteristics that alternately appear and disappear.

An excellent example of a multiple-layer structure is presented in figure 4. Several distinct strata of various vertical thicknesses and temporal behavior can be seen; superimposed is the refractive index and potential temperature profiles obtained from a radiosonde sounding at 1238 PST. The agreement of backscatter zones and refractive index gradients is remarkable. The significance of this layer structure on radar coverage is illustrated in figure 5, in which a transmitter at a height of 100 ft. is assumed. The earth is drawn flat, which gives a straight ray an upward bend. From the ray trace picture we should expect a hole in coverage at a range of around 25 nautical miles. We also see from the ray trace picture that only the layer belonging to the steepest gradient influences the coverage significantly. This layer changed about 80 m in height between 1220 and 1240 PST, which could influence an accurate coverage prediction. This example emphasizes the need for precise and continuous knowledge of the refractive index structure, if accurate coverage predictions are required. Many more examples of layer structures with radiosonde sounding data are presented by Richter and Gossard [1970].

Radar reflectivity is obtained from the recorded A-scope presentation. Examples of quantitative analyses are given by Richter [1969] and Atlas *et al.* [1970]. In general, the maximum reflectivities measured with the FM-CW radar sounder are one to two orders of magnitude higher than those observed with radars having much coarser range resolution.

The radar backscatter properties of water droplets (and, therefore, of rain) are well known. What is not well known is the fine-scale structure of rainfall, and the radar sounder offers a powerful tool for studying it. This possibility has not been fully exploited in San Diego for two reasons. First, it is not associated with the purpose for which the sounder was built; and, second, San Diego has yearly average rainfall of only about 10 in., which is not enough for extensive study. Figure 6 illustrates the capability of the sounder with respect to rain. The upper portion of this figure is the intensity-modulated radar record of a multiple-layer structure with intermittent rain showers starting at 1152 PST and 1222 PST. The echo intensity of the rain overexposed the intensity-modulated record. The A-I record of the time interval from 1151-1158 PST is shown in the lower part of the figure. Structural details of the rain shower are evident. The intensity of the rain with respect to time varies, leading to different "streaks." The slope of the streaks gives the rainfall velocity. The deviations from a constant slope must be due to wind shear. The gap in the record between 1156 to 1157 PST stems from a baseline shift, which is caused by a strong echo (probably bird or airplane) overdriving the radar receiver. Figure 6 is one example of what can be studied with the sounder. Other recordings exist in which rain evaporates before it reaches the ground. The ice-water transition is another problem that could be studied, as the scatter properties of ice crystals are different from those of raindrops.

Perhaps the most exciting observations made with radar sounder concern the detailed structure of wave motions in the troposphere. The presence of such waves has been known [Gossard and Munk, 1954], and they have been recorded by other radars on rare occasions when the size scales and amplitudes were large enough for their much coarser resolution. The present sounder showed that the waves are ubiquitous, and occur with vertical amplitudes of as little as some 10 m. It also revealed details of wave motions that had never been observed before. Many presentations and interpretations of wave motions observed with the sounder have been given [Gossard et al., 1970; Atlas et al., 1970; Gossard and Richter, 1970].

A good example of wave patterns is given in figure 7, which shows the same recording at three different enlargements. Two different types of wave motion are clearly distinguishable: long-period (5 to 10 min. or longer) internal waves, and much shorter-period Kelvin-Helmholtz (K-H) wave instability structures sometimes superimposed on the longer-period waves. The enlargements in the middle and lower part of the figure reveal short-period and small-amplitude K-H waves that show

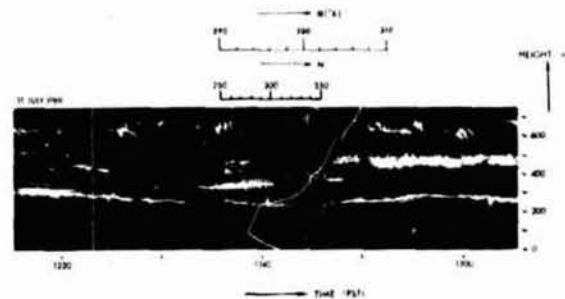


Figure 4.— Multiple layer structure. The dashed curve is the refractive index profile; the solid curve the potential temperature measured with a sounding balloon.

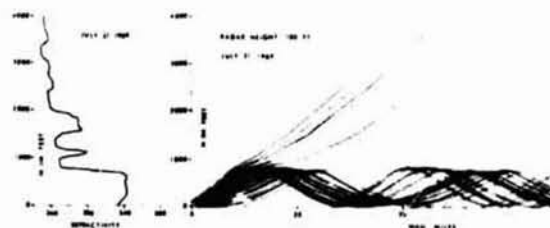


Figure 5.— Ray trace for the refractive index profile and layer structure of figure 4.

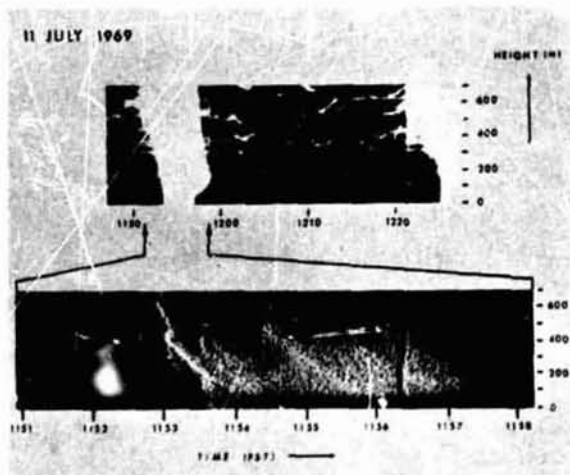


Figure 6.— Structure of rain.

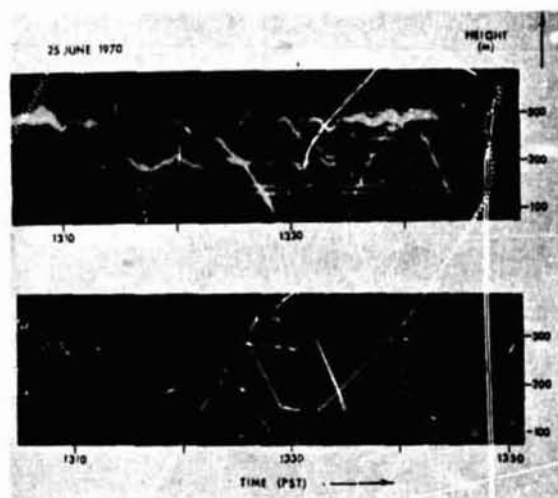


Figure 7.— Wave patterns.

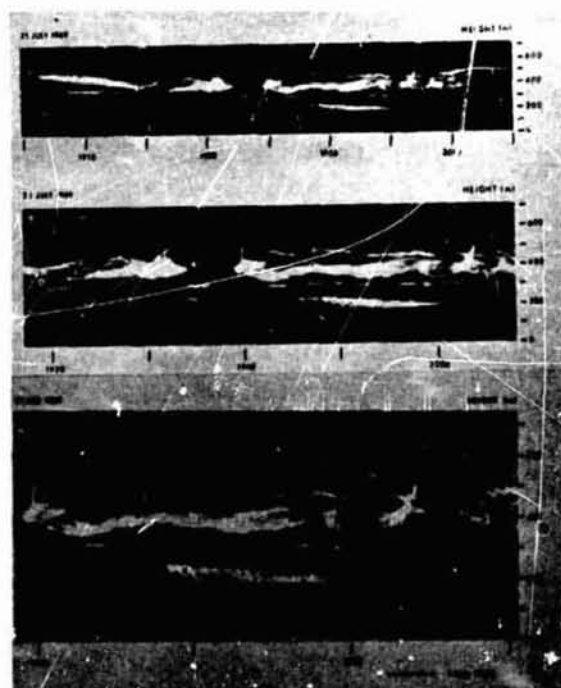


Figure 8.— Breaking wave; upper part, intensity modulation; lower part, A-I presentation.

clear evidence of breaking. A perfect sequence of a K-H wave is shown in figure 8. The wave builds up until it overturns and breaks, and the air masses at the interface become mixed. The lower part shows a section of the same event, but in A-I presentation. This kind of presentation shows particularly clearly the sharp boundaries and the small vertical extent of the zones from which the electromagnetic energy is returned.

Cases in which the radar illuminates the wave structure through considerable depth are of special interest because they can reveal details of the region where true wave motion couples with the instability structures at the critical level. This requires radar returns from multiple layers over considerable depth. Figure 9 shows layers above and below the major wave-train located between 400 and 500 m. The major wave-train perturbs the upper stable layers, but has no apparent effect on the layers below, which tend to consist of independent, breaking wavetrains. The vertical structures between 100 and 170 m are caused by convection in the marine layer below the stable region.

In an effort to examine the fine structure of the scattering layers in more detail, the radar records were recorded on magnetic tape. Since the dynamic range of the magnetic tape is much greater than that of film, various film exposures could be run from the tape recording, and a minimum exposure used for examination of the parts of the layers which would normally be overexposed. When this is done with layers such as those shown in figure 10, an interesting substructure is shown in the lower frame of the figure. It reveals the presence of organized structures (rather than randomly inhomogeneous turbulence) down to the minimum size resolvable by the radar (about 1.5 m in this case). This substructure bears an impressive resemblance to the large instability features such as those shown in figure 8, which have been interpreted as K-H structures due to shear. Gossard *et al.* [1971] have suggested that untrapped internal waves propagating upward or downward through temperature inversion layers can create their own wind shear within the layers, and can also locally modify the temperature gradient. The criterion of dynamic stability (Richardson's number) is thus reduced in certain height regions within the wave structure; and if Richardson's number becomes less than 0.25, there is theoretical reason to believe that

dynamic instability can occur. If gravity wave perturbations occur at the levels of instability (either by propagating from other levels or somehow locally generated), K-H instability should cause those perturbations moving with the local wind speed to grow and eventually break, creating the intense, small-scale turbulent inhomogeneities that are apparently "seen" by the radar. Ambient shear across the stable layer is observed to exist almost 100% of the time in Southern California. Present evidence supports the hypothesis that multiple, thin, dynamically unstable layers are created by untrapped waves (whose wave vector is nearly vertical) imbedded in a medium whose ambient shear is sufficient to locally reduce the Richardson's number below critical.

CONCLUSIONS

The FM-CW radar sounder is a novel tool for the investigation of the structure of the troposphere. It has yielded significant results to date, in particular in the area of radar reflectivity and atmospheric wave motions. The sounder permits continuous monitoring of the vertical refractive index structure in the troposphere with range resolution as high as 1 m. Layer structures are detected routinely. As they may change their height in short time intervals, a continuous knowledge — not only of their presence but also of their exact height — is necessary for accurate radio and radar propagation predictions. The sounder is an excellent tool for the study of the structure of rain. The sounder will continue to play an important role in the field of mesoscale meteorology, in which conventional sounding techniques are extraordinarily difficult.

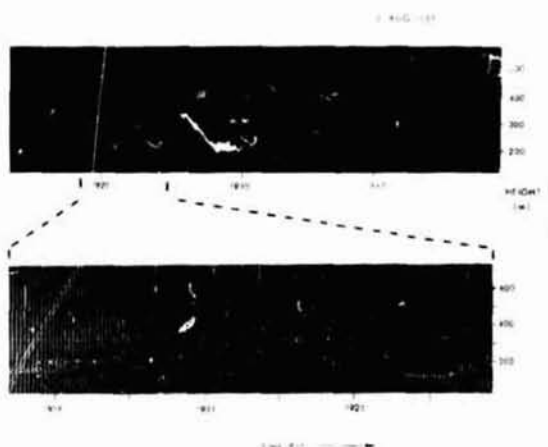


Figure 9.— Wave structures.

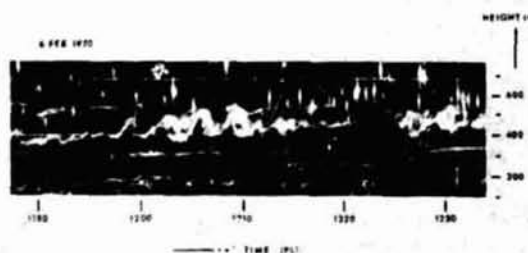


Figure 10.— Top frame shows gravity waves in which radar return is apparently not phase related. Bottom frame shows same record with reduced film exposure and reveals Kelvin/Helmholtz substructure, which apparently accounts for the radar return. Sloping straight lines are captive balloon echo.

DISCUSSION

Wright: In both this and Dr. McAllister's paper I was struck by the instability structure showing two sine waves of about the same period but 180° out of phase — apparently an oscillation between two distinct patterns.

Richter: This has puzzled us for a long time. One should not be deceived, however, and interpret this as a spatial picture. We see the wave motion Doppler shifted by the wind sampled at one point in space.

Unidentified Speaker: Do you have independent wind measurements? Do the waves move with the wind, or do they have a separate propagation velocity?

Richter: They don't necessarily propagate with the wind. Refer to Gossard's paper in JGR, 1970. We do have means of deducing true spatial wavelength with pressure recording measurements on the ground.

Crane: We visualize Kelvin-Helmholtz instabilities that break down and you get turbulence. We see no scattering until the breakdown occurs. But you have the wave well traced before that. Perhaps seeing a different kind of turbulence.

Richter: We talk about turbulence at different scale sizes. The breaking waves themselves consist of smaller scale turbulence and, as I showed, organized substructures. There is much more to be learned. We can relate radar reflectivity to mean-square fluctuations in refractive index, as done notably by Bean in Boulder. But we cannot construct refractive index profiles from the measurements at this time.

Kay: You are working with an FM radar, which results in your integrating over 50 msec. All the returns you see within the resolution element are contributions of temporal integration. Therefore, you are working with turbulence of lower scale than the thickness of layer.

Richter: The integration over 50 msec does not determine our spatial resolution of 1 m. We sample the spectrum of turbulence at half the radar wavelength (5 cm).

Kay: Why do you see the beginning of turbulence?

Richter: Because you see turbulence at a much smaller scale which exists independent of the larger scale turbulence induced by breaking waves.

Bojarski: Therefore we should go to higher resolutions.

Richter: That's right.

Crane: We don't know what we are seeing. Is it turbulence or something else?

LIDAR PROBING OF THE ATMOSPHERE

Some Quantitative Aspects

R. T. H. Collis and E. E. Uthe

Atmospheric Sciences Laboratory
Stanford Research Institute

N73-11595

ABSTRACT

Lidar uses laser energy in radar fashion to observe atmospheric backscattering as a function of range. Because of the short optical and near-optical wavelengths used, very small particles and even the gaseous molecules cause significant scattering. This can complicate the evaluation of the observations by introducing attenuation along the path as a second unknown into the lidar equation. In many cases, however, the observations may be interpreted directly on a *qualitative* basis and show the distribution of particulate matter in clear air or enable the dimensions of visible cloud to be measured accurately. In other cases, particularly where additional data are available, *quantitative* solutions can provide useful information on remote targets such as tenuous smoke clouds or haze layers. Examples of such observations are given, illustrating the computational approach to the evaluation of the volume concentration of natural dust and haze layers in the lower atmosphere and the mass concentration of a smoke plume. In both cases lidar data are related to independently obtained data on the particulate concentrations involved.

INTRODUCTION

Lidar is a technique that uses laser energy in radar fashion for remote probing of the atmosphere. It combines the principles of weather radar with older optical scattering concepts to provide an important new facility for observing particulate concentrations and evaluating atmospheric optical parameters.

The technique has recently been reviewed by Collis [1970], who describes its application in general meteorology and air pollution studies mainly in the lower atmosphere, while Kent and Wright [1970] emphasize its use in upper atmospheric probing. In this paper the problem of deriving quantitative data from lidar observations is discussed and examples are given of successful quantitative interpretation in two actual cases.

BASIC LIDAR CONCEPT

In its basic form, lidar employs a laser simply as a source of pulsed energy. Typically Q-switched ruby or neodymium doped glass systems are used to generate pulses, having peak power of tens of megawatts and durations of 10 to 20 nsec that are directed in highly collimated beams by suitable optical systems. Energy backscattered by the atmosphere is detected by a photomultiplier after collection by a lens or reflector system. The resulting signal is evaluated as a function of time from the transmission of the pulse, in radar fashion, and is typically displayed on an oscilloscope either directly, or after storage in a magnetic disc video recorder. Polaroid or other photographs may be used to record the displayed data.

Taking advantage of the monochromatic nature of the laser energy, narrow band optical filters may be used in the receiver system to minimize "noise" caused by extraneous light. The nature of the received signal P_r is given by

$$P_r(R) = P_t L \beta(R) A R^{-2} \exp \left\{ -2 \int_0^R \sigma(r) dr \right\} \quad (1)$$

where

P_r instantaneous received power

P_t transmitted power at t_0

L effective pulse length (ℓ); $L = c\tau/2$ where c is the velocity of light and τ is pulse duration; it is the range interval from which signals are simultaneously received at time t

β the volume backscattering coefficient of the atmosphere ($\text{sr}^{-1} \ell^{-1}$)

R range [$R = C(t - t_0)/2$] where t_0 is the time of transmission of pulse

σ volume extinction coefficient (ℓ^{-1})

A effective receiver aperture.

The magnitude of β and σ depend on the wavelength of the incident energy, and the number, size, shape, and refractive properties of the illuminated particles per unit volume. These range from the hydrometeors of precipitation, cloud, or fog through the particulate components of the "clear" aerosol, to the molecules of the gaseous atmosphere. The scattering and absorption mechanisms involved have been widely studied. For the moment it will suffice to note that at the wavelengths commonly employed (for example, 0.6943μ for ruby lasers and 1.06μ for neodymium lasers) the energy is readily scattered by particulate material of even very clear atmospheres and typical lidar systems can detect returns from the clear gaseous upper atmosphere.

In relatively clear atmospheres the extinction coefficient is small and variations in backscattering as a function of range can readily be interpreted. For example, the presence of a sudden increase in scattering due to a cloud or haze layer is obvious and unequivocal. Similarly a marked reduction in turbidity is apparent as a decrease of signal intensity at the range of the discontinuity. In strongly scattering conditions, however, marked attenuation occurs and in observing a cloud, for example, the lidar return rapidly diminishes as the penetration of the cloud increases. The evaluation of lidar signals in such circumstances thus presents certain difficulties and imposes the need for care in making qualitative interpretations.

EVALUATION OF THE BASIC LIDAR EQUATION

The atmospheric volume extinction coefficient σ , itself largely due to scattering (although some gaseous and particulate absorption may occur), is in general related to the volume backscattering coefficient β . Where the scattering particles are small compared to wavelength, as is the case with gaseous molecules or the smaller particles of the natural aerosol, Rayleigh scattering applies and the relationship is consistent and $\beta = (1.5/4\pi)\sigma$. However, in so-called "Mie scattering," where particle size is comparable to wavelength, small changes in particle diameter cause marked changes in backscatter cross section. (For example, a change in the diameter of a spherical particle by 1/100 can cause a 20 to 1 variation in its backscatter cross section.) Although substantial averaging occurs over the range of particle sizes present in typical natural aerosols (such as, haze, cloud, or fog), the relation between the volume extinction coefficient and the volume backscattering coefficient is by no means consistent, particularly with the monochromatic energy generated by lasers [Twomey and Ho, 1965].

This uncertainty has important implications, for the lidar equation can only be solved for the optical parameters β and σ if they can be related. Where the relationship remains constant ($\beta/\sigma = k_1$) as a function of range, the lidar signal can be evaluated by solving the lidar equation in a differential form, provided that k_1 is known and a reference value for β or σ is available. This condition applies to atmospheric conditions in which the particle size distribution does not vary substantially and for which values for k_1 can be adduced from knowledge of the particle distribution, or deduced from other lidar measurements (for example, of calibrated targets). Average values of k_1 may be employed for different types of atmospheric targets. In cases where particle size distributions are not constant, as in the case of cloud and various conditions of mist and fog, a power law relationship ($d\ln\beta/d\ln\sigma = k_2$) may be employed to derive at least approximate solutions. Solutions of either form, it should be noted, are very sensitive to errors in the initial assumptions or to errors in measurement.

Additional data in other forms may be useful. For example, the total path extinction or transmittance of a cloud or haze layer may be derived from comparison of observations of the clear air on the near and far side of the layer. The additional information on transmittance may then be used to check and improve the assumptions made regarding the β/σ relationship, as well as to provide reference values. The use of targets of known reflectivity may be used in a somewhat similar way to calibrate lidar performance or derive path transmittance for "visibility" measurement. In the special case of an homogeneous atmosphere (β and σ constant) the lidar equation may be simply solved (for $\sigma = -d\ln P_r R^2 / 2dR$). Similarly, where the optical parameters vary only in the vertical, a solution is possible from a series of two or more lidar observations made at different angles of elevation. In either case, the degree of homogeneity of the atmosphere may be apparent from the lidar observations themselves.

EXAMPLES OF LIDAR OBSERVATIONS OF ATMOSPHERIC CONDITIONS

Observation and Evaluation of Particulate Content of Relatively Clear Air

In this example, sequences of observations made with a neodymium lidar from an aircraft were used to investigate the particulate content in the atmospheric layer below 3 km over the sea in connection with the Barbados Oceanographic and Meteorological Experiment (BOMEX) in 1969 [Uthe and Johnson, 1971].

Figure 1 is a computer-drawn representation of lidar S values. These S values represent range-normalized signal returns in logarithmic form that are corrected for instrumentation transfer anomalies and any pulse-to-pulse lidar performance variations. They thus evaluate in relative terms the atmospheric-dependent parameters of the lidar equation: $\beta \exp -2\int \sigma dr$. Assuming that attenuation may be neglected, a 10x3 km vertical cross section of relative volume concentration of particulate material is shown in absolute terms (fig. 1(a)) and in the form of anomalous departures from a best-fit exponential distribution (fig. 1(b)). This dramatically emphasizes the layer at 1.8 km altitude, which is interpreted as the stream of dust carried by the NE trade winds to the Caribbean area from the Sahara desert. The assumption that attenuation is negligible in such conditions and the relationship of the lidar data to absolute volume concentrations were investigated with the help of particle-size distributions from an independent aircraft sampling program. Using these data and Mie theory expected lidar signal returns were computed: these are shown for three occasions in figure 2.

The dashed lines represent simulated relative lidar returns, ignoring the effect of atmospheric attenuation of the energy pulse. The corresponding unbroken lines represent another set of simulated relative lidar returns obtained after taking into account the attenuation of the laser energy. A lidar altitude of 3 km is assumed. By comparing the dashed and solid profiles, which show only about 1 dB difference at the surface for the two hazy days, it is seen that the neglect of atmospheric attenuation does not significantly affect the return signal profile. Therefore, analysis of particulate mass or volume concentration need not involve the assumptions required for rigid solution of the lidar equation.

The lidar system constant may be inferred by comparing the computed lidar returns with actual lidar returns. The relative signals from three lidar traces each recorded during the aerosol sampling period are shown in figure 2 as

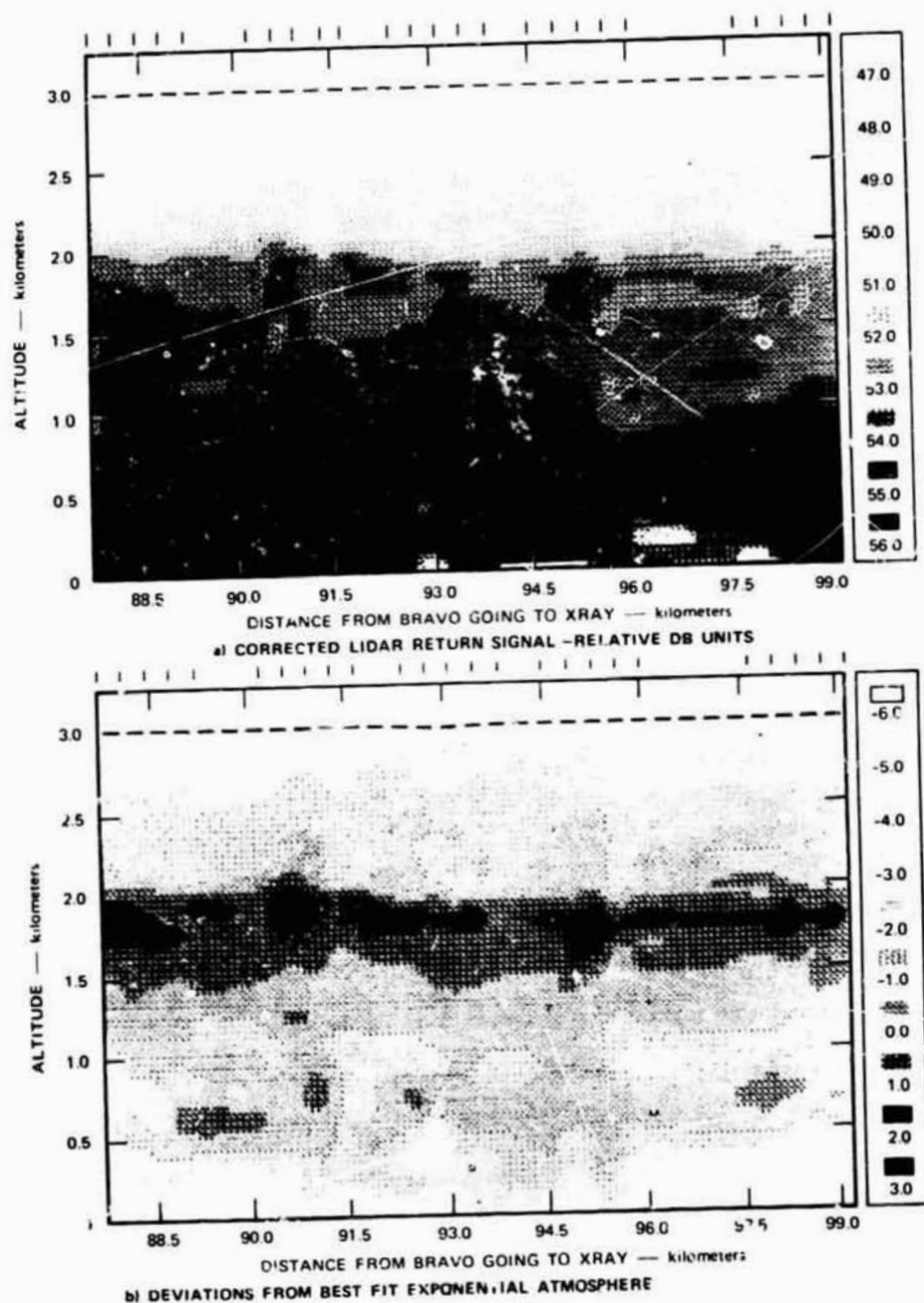


Figure 1.- Digital presentation of lidar observed aerosol structure during BOMEX (30 June 1969, 1053-1056 EDT, Run 30-7)
(Bravo and x-ray are locations in the experimental area)

The location of each observation is indicated by vertical marks along the top lines of the section.

heavy solid lines. The three computed and three observed profiles were first plotted on two separate graphs and the horizontal displacement between these graphs was adjusted for the best overall fit of computed and observed data. The abscissa values represent the observed lidar signals in relative logarithmic units.

Good agreement exists between computed and observed relative lidar return signals in both day-to-day and altitude variations. This result indicates that the lidar characteristics important to the absolute aerosol densities could be inferred.

from the lidar backscatter signatures, subject to errors due to nonlinear variations between particulate density and the volume backscatter coefficient. Figure 3 presents a plot of the particulate volume concentration as a function of lidar return signal. This graph is derived from Mie theory, with the abscissa determined from the best fit between lidar and particle count data. The scatter of the data points is inherent from Mie theory and is a result of insufficient averaging of irregularities between particle size and the particle backscatter cross sections [Twomey and Howell, 1965]. Nevertheless, these investigations show that useful quantitative information on the particulate content of the "clear" air can readily be derived from lidar observations, and, given the density of the particulate phase, the aerosol mass concentration may be inferred.

Measurement of Mass Concentration of a Smoke Plume

In this example the technique was used to infer the total particulate mass per unit length of a plume from the smoke stack of a large power plant [Johnson and Urhe, 1971], illustrating the evaluation of ruby lidar observations when it is necessary to account for attenuation of the laser energy within the target.

The problem may be approached by solving a differential equation in terms of S values derived from the lidar equation (1). Thus

$$\frac{dS}{dR} = 4.34 \frac{1}{\beta} \frac{d\beta}{dR} - 8.7 \sigma \quad (2)$$

where Bouguer's law of attenuation, relating σ to transmittance T ,

$$T(R) = \exp \left[- \int_0^R \sigma(\ell) d\ell \right] \quad (3)$$

has been assumed and $\sigma(\ell)$ is the volume extinction coefficient at a distance ℓ along the path. The solution of Eq. (2) for particulate mass concentration along the laser beam path requires relating the optical parameters σ and β to the particulate mass concentration. For this purpose the ratio of each optical parameter to the particulate mass concentration M (mass/volume) is defined as

$$\xi_e \equiv \frac{\sigma}{M} \quad \xi_b \equiv \frac{\beta}{M} \quad (4)$$

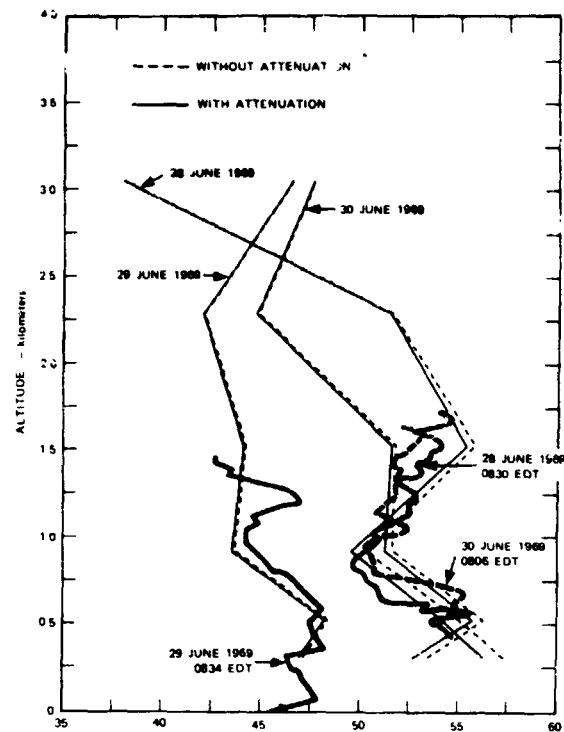


Figure 2.— Derivation of lidar calibration from particle count data. Actual lidar profiles are shown as heavy lines while backscatter profiles evaluated from particle count data are shown as less heavy lines extending to above 3.0 km.

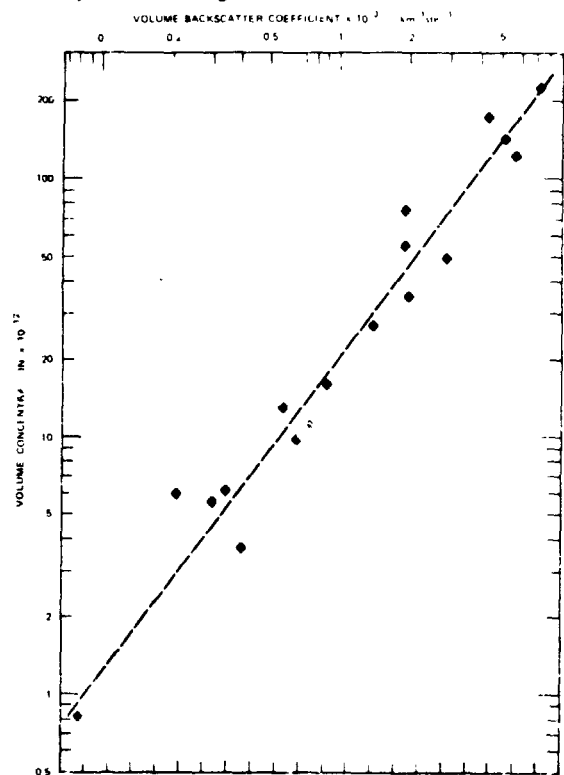


Figure 3.— Lidar return signal and volume backscatter coefficient related to particulate volume concentration.

Substitution of Eq. (4) into the lidar differential equation (Eq. (2)) results in an expression relating mass concentration to the lidar normalized signal return.

$$\frac{dM}{dR} - \frac{1}{4.34} \frac{dS}{dR} M - 2\xi_e M^2 = 0 \quad (5)$$

where ξ_b has been considered not to vary with range. The validity of this assumption will be discussed later. This nonlinear equation may be solved by first employing the linearization transform $\eta \equiv M^{-1}$ with the result

$$M(R) = \exp[C_1 S(R)] \left\{ M_0^{-1} - 2\xi_e \int_0^R \exp[C_1 S(\ell)] d\ell \right\}^{-1} \quad (6)$$

where $C_1 = 1/4.34$. This solution principally assumes that 1) ξ_e and ξ_b are invariant with range and the value of ξ_e is known or can be computed, and 2) a boundary or clear-air particulate mass concentration M_0 is known or can be estimated.

Microscopic analysis of the fly ash effluent from the smoke stack in question clearly showed that the particles were mostly glass spheres, apparently formed from the mineral content of the coal by the intense heat of the combustion process; hence, the use of Mie theory in deriving estimates of ξ_e and ξ_b is justified. To a high degree of accuracy, the volume extinction coefficient at the lidar wavelength of 1.06μ may be considered to be essentially dependent only on the atmospheric particulate matter, and the ξ values may be expressed as:

$$\xi_e = \frac{\int_0^\infty \pi a^2 Q_e(x, m) n(a) da}{\int_0^\infty \frac{4}{3} \pi a^3 \rho n(a) da} \quad (7)$$

$$\xi_b = \frac{\int_0^\infty \pi a^2 Q_b(x, m) n(a) da}{\int_0^\infty \frac{4}{3} \pi a^3 \rho n(a) da} \quad (8)$$

where

- a particle radius
- Q_e extinction efficiency factor
- Q_b backscatter efficiency factor
- x size parameter $= 2\pi a/\lambda$
- m refractive index of particles
- λ wavelength of the laser energy
- n number of particles per unit volume unit radius interval
- ρ density of the particulate matter

To be precise, the parameters ξ_e and ξ_b defined by Eqs. (7) and (8) are exactly equivalent with those defined by Eq. (4) only when multiple scattering effects are absent. This is a good assumption because of the narrow field of view of the lidar system (<1 mrad).

The efficiency factors may be computed from Mie theory for a given particulate matter refractive index, the real part of which in this case was estimated to have a value of approximately 1.5, based on the known mineral composition of coal. The requirement that ξ_e and ξ_b be independent of range is satisfied when the relative particle-size distribution is invariant with range; that is, when

$$n(R,a) = C(R)n_r(a) \quad (9)$$

where $C(R)$ is the number concentration and $n_r(a)$ is the relative size distribution. Also, with this assumption only the relative particle size distribution is required to evaluate the ξ values. However, $n_r(a)$ is rarely completely independent of range, and the question may be asked as to how small changes in the relative size distribution are reflected in the ξ values. Since the ξ values are expressed in terms of integrals over the range of particle sizes, monodisperse (single-size) particles represent the worst case in terms of ξ variations with changes in the particle size. Mie computations of ξ_e and ξ_b are presented in figure 4 for the case of monodisperse particles consisting of material with a density of 2.1 g/cm^3 (as measured for the fly ash in question) and a refractive index of $m=1.5-i0$ for incident radiation of wavelength 1.06μ .

As shown, slight variations in particle size can cause large variations in ξ_b and smaller but significant variations in ξ_e . A real aerosol, however, contains particles of many sizes, so the variations in the ξ curves tend to be smoothed out when a weighted integral for all sizes present is computed. Relative particle-size distribution data independently obtained by sampling from an aircraft were made available to us, along with information about sampling volumes and total particle counts. Some of these sets of data were reduced to absolute size distributions, to which a Junge-type model was fitted (that is, of the form $n = a_0 a^{-\gamma}$, where a_0 and γ are fitted constants, and n and a are as defined earlier). However, there is some uncertainty about the accuracy of the sampling volume values used to compute the absolute distributions, since direct conversion of some of the clear-air measurements to mass concentrations gave unrealistically high values ($M_0 > 1000 \mu\text{g/m}^3$) of concentration in the clear air upwind of the stack. Since the relative size distributions are sufficient for use with the lidar data and also are felt to be more accurate, they have been used exclusively in our analyses. The evaluation of the solution parameters ξ_e and ξ_b from these particulate data (24 samples taken on several days at various times and at various positions within the plume) gave values ranging from 0.2 to $1.0 \text{ m}^2/\text{g}$ and from 0.014 to $0.064 \text{ m}^2/\text{g}\cdot\text{sr}$, respectively. The corresponding overall means were $0.55 \text{ m}^2/\text{g}$ and $0.039 \text{ m}^2/\text{g}\cdot\text{sr}$.

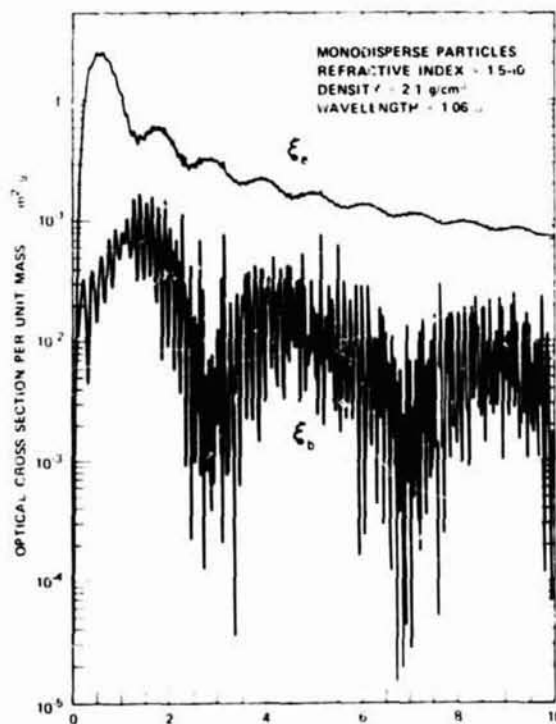


Figure 4.— Optical cross sections per unit mass as a function of particle size computed from Mie theory for fly ash.

The lidar observations and particle sampling of the plume were best coordinated for cross section 64, one of the series of vertical sections made through the plume by scanning the lidar in elevation. This cross section was oriented approximately cross wind, and this case was chosen to apply the mass concentration solution discussed previously. The boundary parameter M_0 was taken from an estimate of visibility (7 to 10 miles), which gives $M_0 = 100 \mu\text{g/m}^3$, using the relationship of Noll *et al.* [1968]. Computation of ξ_e with the aid of the particle-size measurements gave a value of $0.25 \text{ m}^2/\text{g}$.

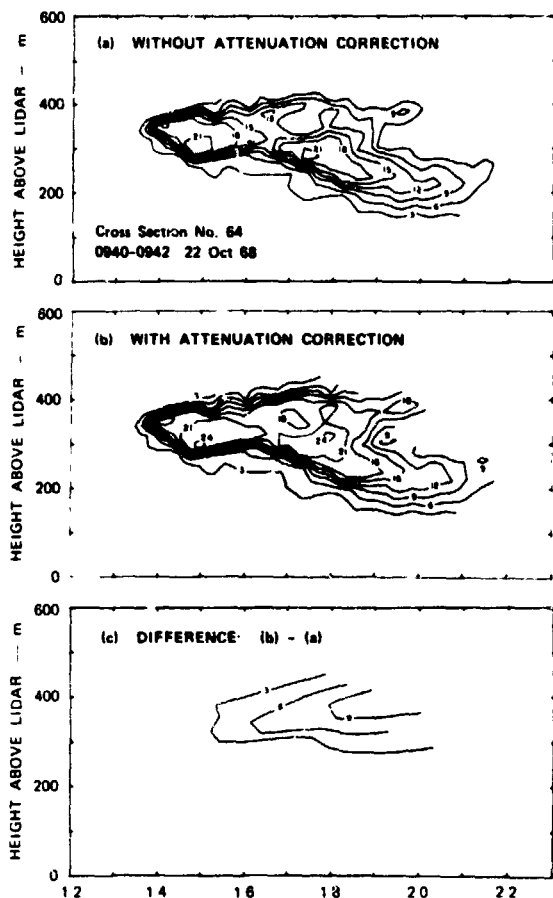


Figure 5. - Lidar observations of plume cross section 64 in relative logarithmic units.
(a) Without attenuation correction
(b) With attenuation correction
(c) Attenuation correction

With this information, the lidar S function cross section (fig. 5(a)) was converted by using Eq. (6) to an attenuation-corrected cross section of absolute mass concentration (fig. 5(b)). Figure 5(c) represents the results of the graphical subtraction of (b) from (a), and thus shows contours of mass concentration (in logarithmic units) added to the further side of the plume by the attenuation correction. It may be noted that the maximum correction is 9 dB, which means that the mass concentration in that region of the uncorrected cross section in figure 5(a) was almost 10 times too low. Although the attenuation correction does change concentrations significantly in one part of the cross section, the general shape and structure of the plume remain the same.

A planimeter was used to integrate the total mass per unit plume length represented by the cross sections in figure 5(a) and 5(b). The results were as follows: uncorrected cross section 5(a), 350 g/m; corrected 5(b), 680 g/m. The attenuation correction thus was significant in terms of the mass content. These values were compared with an estimate of the particulate emission rate from the stack. Using observed data on plant operation and standard conversion techniques gave a plume lengthwise mass content of 875 g/m, compared with the 680 g/m measured by the lidar after allowance for attenuation. Although further evidence will be necessary from other comparisons, the close agreement of this initial example does indicate that the lidar technique may be capable of useful quantitative measurements of mass concentration.

CONCLUSIONS AND OUTLOOK FOR THE FUTURE

The lidar technique is a unique method of observing atmospheric structure and opens up new ways of deriving intelligence of atmospheric conditions and processes. This intelligence can be a vital input into many operational activities in meteorology and atmospheric physics such as aviation and air pollution, to name but two obviously important areas. In research, the value of lidar is no less great. Even with the relatively modest equipment used to date, lidar has provided a new insight into a number of atmospheric phenomena and revealed shortcomings in our current understanding of certain processes (for example, in turbulent diffusion from high smoke stacks).

In certain applications its role will be to make *quantitative* observations of atmospheric characteristics. In some cases, however, it may well be that the difficulties of reducing and interpreting the observation in absolute terms will limit the use of lidar to *semiquantitative* observations. The remote detection and mapping of structural features of the atmosphere by reference to *relative* variations in its optical characteristics is a most valuable capability, however, with obvious utility in many areas of atmospheric research and operational meteorology.

DISCUSSION

Unidentified Speaker: Have you evaluated the role of multiple scattering in your signal?

Uthe: Yes. Multiple scattering normally plays very little role in lidar scattering because we are using milliradian beamwidths and receiver field-of-views. The only important multiple-scattering effect is the energy scattered back into the forward directed beam, and this can be accounted for by using the aerosol forward-scattering coefficient.

ACTIVE ATMOSPHERIC SOUNDING

BIBLIOGRAPHY

- Andermo, I. Fornaeus, L., Svensson, I., and Wiborg, G., Laser-Molnhöjdmätare: Beskrivning och Matningar Gjorda T.O.M. September, 1964, FOA 2 Rapport, Försvarets Forskningsanstalt Stockholm, 1965.
- Army, Dept. of, Artillery Meteorology, Field Manual FM6-15, Headquarters, Washington, D.C., 1962.
- Ashwell, G. E., and Lane, J. A., Simultaneous Lidar and Refractometer Soundings of Layer Structure in the Troposphere, Nature, 222, 464-466, 1969.
- Atlas, D., Radar Detection of the Sea Breeze, J. Meteorol., 17, 244-258, June 1960.
- Atlas, D., Indirect Probing Techniques, Bull. Amer. Meteorol. Soc., 43, 457-466, 1962.
- Atlas, D., Advances in Radar Meteorology, Advan. Geophys., 10, 317-478, 1964.
- Atlas, D., and Hardy, K. R., Radar Analysis of the Clear Atmosphere: Angels, in Proc. 15th General Assembly of URSI, 401-469, Int. Sci. Radio Union, Munich, Germany, 1966.
- Atlas, D., Hardy, K. R., and Naito, K., Optimizing the Radar Detection of Clear Air Turbulence, J. Appl. Meteorol., 5, 450-460, 1966.
- Atlas, D., Harris, F. I., and Richter, J. H., Measurement of Point Target Speeds with Incoherent Non-Tracking Radar: Insect Speeds in Atmospheric Waves, J. Geophys. Res., 75, 7588-7595, 1970.
- Atlas, D., Metcalf, J. I., Richter, J. H., and Gossard, E. E., The Birth of CAT and Microscale Turbulence, J. Atmos. Sci., 27, 903-913, 1970.
- Baerg, W., and Schwarz, W. H., Measurements of the Scattering of Sound From Turbulence, J. Acoust. Soc. Amer., 39, 1125-1132, June 1966.
- Barrett, E. W., and Ben-Dov, O., Application of Lidar to Air Pollution Measurements, J. Appl. Meteorol., 6, 500-515, 1967.
- Batchelor, G. K., Wave Scattering Due to Turbulence, Symp. Naval Hydrodynamics, F. S. Sherman, ed., NAS-NRC Publ. 515, 409-430, 1957.
- Battan, L. J., Radar Meteorology, University of Chicago Press, Chicago, 1959.
- Bean, B. R., and McAllister, L. G., A Comparison of In-Situ and Remote Sensing of the Boundary Layer, Remote Sensing Techniques in Geophysics, XV IUGG General Assembly, Moscow, 1971.
- Becker, R., Shockwave and Detonation, Z. Phys., 8, 321-362, 1921.
- Bellamy, J. C., and Bosshart R. F., Investigation of the Feasibility of Acoustic Soundings of the Atmosphere, Cook Res. Labs, Skokie, Illinois, Final Report, 1958.

Beran, D. W., Project EAR (Environmental Acoustic Research) Vol. 1 and 2, University of Melbourne, Meteorology Dept. Pub. No. 15 and 17, 1970.

Berto'otti, M., Muzil, L. and Sette, D., On the Possibility of Measuring Optical Visibility by Using a Ruby Laser, Appl. Opt., 8, 117-120, 1969.

Bird, L. G., and Rider, N. E., The Laser Cloud-Base Recorder, Meteorol. Mag., (London), 97, 107-115, 1968.

Booker, H. G., and Gordon, W. E., A Theory of Radio Scattering in the Troposphere, Proc. IRE, 38, 401-412, 1950.

Bureau, R., Altimétrie des Nuages per Impulsions Luminenses, Météorologie, (Paris), 3, 292-301, 1946.

Clemesha, B. R., Kent, G. S., and Wright, R. W. H., A Laser Radar for Atmospheric Studies, J. Appl. Meteorol., 6, 386-395, 1967.

Cohen, A., Neumann, J., and Low, W., An Experimental Determination of the Polarization of Scattered Laser Light by Atmospheric Air, J. Appl. Meteorol., 8, 952-954, 1969.

Collis, R. T. H., Lidar Observations of Cloud, Science, 149, 978-981, 1965.

Collis, R. T. H., Lidar-A New Atmospheric Probe, Quart. J. Roy. Meteorol. Soc., 92, 220-230, 1966.

Collis, R. T. H., Lidar Observations of Atmospheric Motion in Forest Valleys, Bull. Amer. Meteorol. Soc., 49, 918-922, 1968.

Collis, R. T. H., Lidar Observations of Cirrus Cloud, Proc. 13th Weather Radar Conf., J. S. Marshall, ed., 530-534, Amer. Meteorol. Soc., Boston, 1968.

Collis, R. T. H., Lidar, Appl. Opt., 9, 1782-1788, 1970.

Collis, R. T. H., Fernald, F. G., and Alder, J., Lidar Observations of Sierra Wave Conditions, J. Appl. Meteorol., 7, 227-233, 1968.

Collis, R. T. H., Fernald, F. G., and Ligda, M. G. H., Laser Radar Echoes from a Stratified Clear Atmosphere, Nature, 203, 1274-1275, 1964.

Collis, R. T. H., and Ligda, M. G. H., Laser Radar Echoes from the Clear Atmosphere, Nature, 203, 508, 1964.

Collis, R.T.H., Viezee, W., Uthe, E. E., and Oblanas, J., Visibility Measurement for Aircraft Landing Operations, SRI Final Report, AFCRL Contract F 19628-70-C-0083 and FAA Contract DOT-FA 70WA1-178, 1970.

- Cook, R. K., Strange Sounds in the Atmosphere, Sound, 1, 12-16, 1962 2 references, Cook, R. K. and Young, J. M., Sound, 1, 25-33, 1962.
- Cooney, J., Measurements on the Raman Component of Laser Atmospheric Back Scatter, Appl. Phys. Lett., 12, 40-41, 1968.
- Cooney, J., Orr, J., and Tomasetti, C., Measurements Separating the Gaseous and Aerosol Components of Laser Atmospheric Backscatter, Nature, 224, 1098-1099, 1969.
- Cooney, J., Remote Measurements of Atmospheric Water Vapor Profiles Using the Raman Component of Laser Backscatter, J. Appl. Meteorol., 9, 182-184, 1970.
- Cox, E. F., Upper Atmosphere Temperatures From Helgoland Big Bang, J. Meteorol., 6, 300, 1949.
- Crane, R. K., Simultaneous Radar and Radiometer Measurements of Rain Shower Structure, Technical Note 1968-33, Lincoln Laboratory, MIT, Sept. 1968.
- Crane, R. K., A Comparison Between Monostatic and Bistatic Scattering From Rain and Thin Turbulent Layers, Technical Note 1970-29, Lincoln Laboratory, MIT, Oct. 1970a.
- Crane, R. K., Measurement of Clear Air Turbulence in the Lower Stratosphere Using the Millstone Hill L-Band Radar, Proc. 14th Radar Meteorol. Conf., 101-106, Amer. Meteorol. Soc., Boston, 1970b.
- Crane, R. K., Description of the Avon-to-Westford Experiment, Tech. Rep. 483, Lincoln Laboratory, MIT, April 1971.
- Crav, A. P., Stratosphere Winds and Temperatures From Acoustical Propagation Studies, J. Meteorol., 7 233, 1950.
- Crooks, W. M., Hoblit, F. M., and Prophet, D. T., Project Hicat, An Investigation of High Altitude Clear Air Turbulence, Vol. 1, Air Force Flight Dynamics Lab. Tech. Rept. AFFDL-TR-67-123, WPAFB 1967.
- Curcio, J. A., and Knestrick, G. L., Correlation of Atmospheric Transmission with Backscattering, J. Opt. Soc. Amer., 48, 686-689, 1958.
- Davis, P. A., The Analysis of Lidar Signatures of Cirrus Clouds, Appl. Opt., 8, 2099-2102, 1969.
- Davis, P. A., Airborne Ruby-Lidar and Radiometric Measurements of Cirrus and Haze During Bomex, SRI Final Report ESSA Contract E-263-68, 1970.
- deBelatini, P. C. M., Inadequacy of Scatter Mechanisms in Tropospheric Radio Propagation, Nature, 184, 1558-1559, 1959.
- Deirmendjian, D., Electromagnetic Scattering on Spherical Polydispersions, Elsevier, New York, 1969.
- Elterman, L., Wexler, R., and Chang, D. T., Features of Tropospheric and Stratospheric Dust, Appl. Opt., 8, 893-903, 1968.

- Evans, W. E., Wiegman, E. J., Viezee, W., and Ligda, M.G.H., Performance Specifications for a Meteorological Satellite Lidar, SRI Final Report, NASA Contract NASr-49(22), 1966.
- Eyges, L. J., Proposal for Improving Tropospheric Propagation, MIT, Lincoln Lab., Lexington, Mass., Lincoln Lab. Rep. TR-151, 1957.
- Fehlhaber, L., and Grosskopf, J., An Investigation of the Structure of the Troposphere by Means of a Vertically Pointed Radar, Nachr. Z. Commun. J., 4, 185-190, 1965.
- Fenn, R. W., Correlation Between Atmospheric Backscattering and Meteorological Visual Range, Appl. Opt., 5, 293, 1966.
- Fetter, R. W., Remote Measurement of Wind Velocity by the Electromagnetic Acoustic Probe, II. Experimental System, Conf. Proc., 5th Nat. Conv. on Military Electronics, Wash. D.C., Midwest Research Institute, Report No. 420, 54-59, 1961.
- Fiocco, G., and DeWolf, J. B., Frequency Spectrum of Laser Echoes From Atmospheric Constituents and Determination of Aerosol Content of Air, J. Atmos. Sci., 25, 488-496, 1968.
- Fiocco, G., and Smullin, L. O., Detection of Scattering Layers in the Upper Atmosphere (60-140 km) by Optical Radar, Nature, 199, 1275-1276, 1963.
- Ford, G. W., and Meecham, W. C., Scattering of Sound by Isotropic Turbulence of Large Reynolds Number, J. Acoust. Soc. Amer., 32, 1668-1672, 1960.
- Fox, H. L., Meteorological Techniques for Sound Ranging: Conceptual Basis, Tech. Rep. ECOM-00151-4, U.S. Army Electronics Command, Fort Monmouth, N.J., 1966.
- Fox, H. L., and Wiener, F. M., A Simple Algorithm for Determining the Location of a Focus in Atmospheric Sound Propagation, J. Acoust. Soc. Amer., 39, 1260, 1966.
- Fox, H. L., and Sherr, P. E., Meteorological Techniques for Sound Ranging: Results of Computer Studies, Tech. Rept. ECOM-00151-F, U.S. Army Electronics Command, Fort Monmouth, N.J., 1967.
- Fox, H. L., Noiseux, D. U., Starr, E. A., Chakravani, K. L., and Robinson, S. P., Signal Processing Techniques for Sound Ranging: Preliminary Studies, Tech. Rept. ECOM-0378-1, U.S. Army Electronics Command, Fort Monmouth, N.J., 1967.
- Friedland, S. S., Katzenstein, J., and Zatzick, M. A., Pulsed Searchlighting the Atmosphere, J. Geophys. Res., 61, 415-434, 1956.
- Friend, A. W., Theory and Practice of Tropospheric Sounding by Radar, Proc. Inst. Radio Engr., 37, 116-137, 1949.
- Gething, J. T., and Jenssen, D., Measurement of Temperature and Humidity by Acoustic Echo Sounding, Nature (Phys. Sci.), 231, 198-200, June 28, 1971.
- Gilman, G. W., Coxhead, H. B., and Willis, F. H., Reflection of Sound Signals in the Troposphere, J. Acoust. Soc. Amer., 18, 274-283, Oct. 1946.

- Golitsyn, G. S., Gurvich, A. S., and Tatarskii, V. I., Investigation of the Frequency Spectra of Amplitude and Phase Difference Fluctuations of Sound Waves in a Turbulent Atmosphere, Soviet Acoust., 6, 185-194, 1960.
- Gordon, W. E., Radio Scattering from a Turbulent Elevated Layer, Proc. Conf. Radio Meteorol., Univ. Texas, Nov. 9-12, 1953.
- Gossard, E. E., Reflection of Microwaves by a Refractive Layer Perturbed by Air Waves, IRE Trans. Antennas Propagat., APID 317-325, 1962.
- Gossard, E. E., Vertical Flux of Energy into the Lower Atmosphere From Internal Gravity Waves Generated in the Troposphere, J. Geophys. Res., 67, 745-757, 1962.
- Gossard, E. E., Jensen, D. R., and Richter, J. H., An Analytical Study of Tropospheric Structure Seen by High Resolution Radar, J. Atmos. Sci., 28, 794-807, 1971.
- Gossard, E. E., and Munk, W., On Gravity Waves in the Atmosphere, J. Meteorol., 11, 259-269, 1954.
- Gossard, E. E., and Richter, J. H., The Shape of Internal Waves of Finite Amplitude From High Resolution Radar Sounding of the Lower Atmosphere, J. Atmos. Sci., 27, 971-973, 1970.
- Gossard, E. E., Richter, J. H., and Atlas, D., Internal Waves in the Atmosphere From High-Resolution Radar Measurements, J. Geophys. Res., 75, 3523-3536, June 20, 1970.
- Goyer, G., and Watson, R., The Laser and its Application to Meteorology, Bull. Amer. Meteorol. Soc., 44, 564-570, 1963.
- Grams, G. W., Laser Radar Studies of the Atmospheric Aerosol Above 50 km, J. Atmos. Terr. Phys., 32, 729-736, 1970.
- Guttenberg, B., The Velocity of Sound Waves and the Temperature in the Stratosphere in Southern California, Bull. Am. Meteorol. Soc., 20, 192, 1939.
- Hall, F. F., and Ageno, H. Y., Absolute Calibration of a Laser System for Atmospheric Probing, Appl. Opt., 9, 1820-1824, 1970.
- Hamilton, P. M., The Use of Lidar in Air Pollution Studies, Air & Water Pollut. Int. J., 10, 427-434, 1966.
- Hamilton, P. M., Lidar Measurement of Backscatter and Attenuation of Atmospheric Aerosol, Atmos. Environ., 3, 221-223, 1968.
- Hamilton, P. M., The Application of a Pulsed-Light Range-Finder (Lidar) to the Study of Chimney Plumes, Phil. Trans. Roy. Soc. London, A265, 153-183, 1969.

- Hardy, K. R., Atlas, D., and Glover, K. H., Multiwavelength Backscatter From the Clear Atmosphere, J. Geophys. Res., 71, 1537-1557, 1966.
- Hardy, K. R., and Katz, I., Probing the Clear Atmosphere with High Power High Resolution Radars, Proc. IEEE, 57, 468-480, 1969.
- Harris, C. M., Absorption of Sound in Air in the Audio-Frequency Range, J. Acoust. Soc. Amer., 35, 11, 1963.
- Harris, C. M., Absorption of Sound in Air Versus Humidity and Temperature, J. Acoust. Soc. Amer., 40, 148-159, 1966.
- Hicks, J. J., Radar Observations of Gravitational Waves in a Clear Atmosphere, Proc. 13th Radar Meteorol. Conf., 258-261, Amer. Meteorol. Soc., Boston, Mass., 1968.
- Hicks, J. J., Radar Observations of a Gravitational Wave in Clear Air Near the Tropopause Associated with CAT, J. Appl. Meteorol., 8, 627-633, 1969.
- Hicks, J. J., and Angell, J. K., Radar Observations of Breaking Gravitational Waves in the Visually Clear Atmosphere, J. Appl. Meteorol., 7, 114-121, 1968.
- Hines, C. O., Internal Atmospheric Gravity Waves at Ionospheric Heights, Can. J. Phys., 38, 1441-1481, 1960.
- Hirono, M., On the Observation of the Upper Atmospheric Constituents by Laser Beams, J. Radio. Res. Lab., Tokyo, 11, 251-271, 1964.
- Horman, M. H., Measurement of Atmospheric Transmissivity Using Backscattered Light from a Pulsed Light Beam, J. Opt. Soc. Amer., 51, 681-691, 1961.
- Huffaker, R. M., Jelalian, A. V., and Thomson, J. A. L., Laser-Doppler System for Detection of Aircraft Trailing Vortices, Proc. IEEE, 58, 322-326, 1970.
- Jacobowitz, H., and Wark, D. Q., On Cloud-Top Determination from Gemini-5, J. Atmos. Sci., 24, 63-69, 1967.
- Johnson, C. P., and Hale, F. E., Abnormal Sound Propagation Over the Southwestern United States, J. Acoust. Soc. Amer., 25, 642, 1953.
- Johnson, W. B., Lidar Applications in Air Pollution Studies, J. Air Pollut. Contr. Assoc., 19, 176-189, 1969.
- Johnson, W. B., Lidar Observations of Diffusion and Rise of Stack Plumes, J. Appl. Meteorol., 8, 443-449, 1969.
- Johnson, W. B., and Uthe, E. E., Lidar Study of the Keystone Stack Plume, Atmos. Environ., 5, 703-724, 1971.

- Kallistratova, M. A., An Experimental Investigation into the Scattering of Sound in a Turbulent Atmosphere, Dokl. Akad. Nauk. SSSR, **125**, 69-72, 1959.
- Kallistratova, M. A., Experimental Investigation of Sound Wave Scattering in the Atmosphere, Tr. Inst. Fiz. Atmos., Atmos. Turbulentnost, **4**, 203-256, 1961.
- Kallistratova, M. A. and Tatarski, V. I., Accounting for Wind Turbulence in the Calculation of Sound Scattering in the Atmosphere, Akust. Zh., **6**, 503-505, October-December 1960, (Soviet Phys.-Acoustics, **6**, 503-505, 1960).
- Kato, Y., Mori, Y., and Inaba, H., On the Possibility of the Upper Atmospheric Sounding by Optical Radar, Rep. of Ionosphere and Space Res. in Japan XVII, 103-108, 1964.
- Kelton, G., and Bricout, P., Wind Velocity Measurements Using Sonic Techniques, Bull. Amer. Meteorol. Soc., **45**, 571-580, 1964.
- Kent, G. S., and Wright, R. W., A Review of Laser Radar Measurements of Atmospheric Properties, J. Atmos. Terr. Phys., **32**, 917-943, 1970.
- Konrad, T. G., The Dynamics of the Convective Process in Clear Air as Seen by Radar, J. Atmos. Sci., **27**, 1138-1147, 1970.
- Kraichnan, R. H., The Scattering of Sound in a Turbulent Medium, J. Acoust. Soc. Amer., **25**, 1096-1104, Nov. 1953.
- Laird, D. T., Spherical Sound Waves of Finite Amplitude, Ph.D Thesis, Penn. State Univ., 1955.
- Lane, J. A., Ashwell, G. E., and Dagnall, A., Some Results of Lidar Probing of the Troposphere, Atmos. Environ., **5**, 49-54, 1971.
- Lauvstad, V., and Tjøtta, S., Problem of Sound Scattered by Sound, J. Acoust. Soc. Am., **34**, 1045-1050, 1962.
- Leonard, D. A., Observation of Raman Scattering from the Atmosphere Using a Pulsed Nitrogen Ultraviolet Laser, Nature, **216**, 142, 1967.
- Ligda, M. G. H., Meteorological Observations with Pulsed Laser Radar, Proc. 1st Conf. on Laser Technology, G. Adelman and T. B. Dowd, eds., (Office of Naval Research, Boston Office, 1964), **1**, 63, 1963.
- Lighthill, M. G., On the Energy Scattered from the Interaction of Turbulence with Sound or Shock Waves, Proc. Cambridge Phil. Soc., **49**, 531-555, 1953.
- Little, C. G., Acoustic Methods for the Remote Probing of the Atmosphere, Proc. IEEE, **57**, 571-578, 1969.
- McAllister, L. G., Acoustic Sounding of the Lower Troposphere, J. Atmos. Terr. Phys., **30**, 1439-1440, 1968.

- McAllister, L. G., Pollard, J. R., Mahoney, A. R., and Shaw, P. J. R., Acoustic Sounding—A New Approach to the Study of Atmospheric Structure, Proc. IEEE, 57, 579–587, 1969.
- McAllister, L. G., and Pollard, J. R., Acoustic Sounding of the Lower Atmosphere, Proc. Sixth Int. Symp. Remote Sensing of Environment, Ann Arbor, Michigan, 1969.
- McCormick, R. A., and Baulch, D. M., The Variation with Height of Dust Loading Over a City as Determined from Atmospheric Turbidity, J. Air Pollut. Contr. Assoc., 12, 492–496, 1962.
- Melfi, S. H., Lawrence, J. D., and McCormick, M. P., Observation of Raman Scattering by Water Vapor in the Atmosphere, App. Phys. Lett., 15, 295–297, 1969.
- Middleton, W. E. K., and Spilhaus, A. F., Meteorological Instruments, p. 208, Univ. Toronto Press, Toronto, 1953.
- Monin, A. S., Characteristics of the Scattering of Sound in a Turbulent Atmosphere, Akust. Zh., 7, 457–461, 1961 (Sov. Phys. Acoustics, 7, 370–373, 1962).
- Munick, R. J., Turbulent Backscatter of Light, J. Opt. Soc. Amer., 55, 893–894, 1965.
- Nell, K. E., Mueller, P. E., and Imada, M., Visibility and Aerosol Concentration in Urban Air, Atmos. Environ., 2, 465–475, 1968.
- Northend, C. A., Honey, R. C., and Evans, W. E., Laser Radar (Lidar) for Meteorological Observations, Rev. Sci. Instr., 37, 393–400, 1966.
- Nyborg, W. L., and Mintzer, D., Review of Sound Propagation in the Lower Atmosphere, WADC Tech. Report 54-602, May 1955.
- Perkins, B., Jr., and Jackson, W. F., Handbook for Prediction of Air Blast Focusing, BRL Report 1240, Ballistic Research Laboratories, Aberdeen Proving Ground, Md., 1964.
- Perlat, A., and Tett, M., Mesures en Météorologie, Gauthier-Villars, Paris, 1961.
- Pilipowsky, S., Weinman, J. A., Clemesha, B. R., Kent, G. S., and Wright, R. W., Investigation of the Stratospheric Aerosol by Infrared and Lidar Techniques, J. Geophys. Res., 73, 7553–7560, 1968.
- Pridmore-Brown, D. C., and Ingard, U., Sound Propagation into the Shadow Zone in a Temperature-Stratified Atmosphere Above a Plane Boundary, J. Acoust. Soc. Amer., 27, 36, 1955.
- Richardson, J. M., and Kennedy, W. B., Atmospheric Winds and Temperatures to 50-kilometers Altitude as Determined by Acoustical Propagation Studies, J. Acoust. Soc. Amer., 24, 731, 1955.
- Richter, J. H., High Resolution Tropospheric Radar Sounder, Radio Sci., 4, 1261–1268, 1969.

Richter, J. H., and Gossard, E. E., Lower Tropospheric Structure as Seen by a High-Resolution Radar, Naval Electronics Lab Rep. NELC TR 1718, June 26, 1970.

Rothwell, P., Sound Propagation in the Lower Atmosphere, J. Acoust. Soc. Amer., 28, 656, 1956.

Rudnick, I., On the Attenuation of High Amplitude Waves of Stable Form Propagated in Horns, J. Acoust. Soc. Amer., 30, 339, 1958.

Sandford, M. C. W., and Gibson, A. J., Laser Radar Measurements of the Atmospheric Sodium Layer, J. Atmos. Terr. Phys., 32, 1423-1430, 1970.

Saxton, J. A., Lane, J. A., Meadows, R. W., and Mathews, P. A., Layer Structure of the Troposphere, Proc. IEEE, London, 111, 275-283, 1964.

Schotland, R. M., Some Aspects of Remote Atmospheric Sensing by Laser Radar, in Vol. 2, Final Report of the Panel on Remote Atmospheric Probing, D. Atlas, ed., Comm. Atmos. Sci., Nat. Acad. Sci., NRC, 1969.

Schotland, R. M., Nathan, A. M., Chermack, E. A., and Uthe, E. E., Optical Sounding, Tech. Rept. 2, New York Univ., Contract DA-36-039-SC87299, U.S. Army, 1962.

Schuster, B. G., Detection of Tropospheric and Stratospheric Aerosol Layers by Optical Radar (Lidar), J. Geophys. Res., 75, 3123-3132, 1970.

Sekera, Z., Helmholtz Waves in a Linear Temperature Field with Vertical Wind Shear, J. Meteorol., 5, 93-102, 1948.

Smith, P. L., Jr., Remote Measurement of Wind Velocity by the Electromagnetic Acoustic Probe, I. System Analysis, Conf. Proc. 5th Nat. Conv. on Military Electronics, Wash., D.C., Midwest Res. Inst., Rep. No. 419, 48-53, 1961.

Stroud, W. G., Nordberg, W., and Walsh, J. R., Atmospheric Temperatures and Winds Between 30 and 80 km, J. Geophys. Res., 61, 45-56, January 1956.

Tatarski, V. I., Wave Propagation in a Turbulent Medium, (R. A. Silverman, transl.), New York: McGraw-Hill, 1961.

Tolstoy, I., and Herron, T. J., A Model for Atmospheric Pressure Fluctuations in the Meso Scale Range, J. Atmos. Sci., 26, 270-273, 1969.

Twomey, S., and Howell, H. B., The Relative Merit of White and Monochromatic Light for the Determination of Visibility by Backscattering Measurement, Appl. Opt., 4, 501-506, 1965.

Uthe, E. E., and Johnson, W. B., Lidar Observations of the Lower Tropospheric Aerosol Structure During Bomex, SRI Final Report, AEC Contract AT(04-3)-115, 1971.

Van de Hulst, H. G., Light Scattering by Small Particles, Wiley, New York, 1957.

Viezee, W., Uthe, E. E., and Collis, R. T. H., Lidar Observations of Airfield Approach Conditions—An Exploratory Study, J. Appl. Meteorol., 8, 274–283, 1969.

Viezee, W., Collis, R. T. H., and Oblanas, J., Lidar Observations in Relation to the Atmospheric Winds Aloft, J. Appl. Meteorol., 9, 916–920, 1969.

Warner, J., and Telford, J. W., Convection Below Cloud Base, J. Atmos. Sci., 24, 374–382, 1967.

Weinman, J. A., and Ueyoshi, K., The Effect of the Phase Function at Forward Angles on Light Pulse Scattered Backward from a Thin Turbid Medium, J. Atmos. Sci., 26, 600–603, 1969.

Whipple, F. J. W., The Propagation of Sound to Great Distances, Quart. J. Roy. Meteorol. Soc., 61, 285, 1935.

Wyman, P. W., Laser Radar Eye Hazard Considerations, Appl. Opt., 8, 383–392, 1969.

3. RADIO-OCCULTATION MEASUREMENTS

This chapter which treats both active atmospheric and ionospheric sounding, forms a bridge between Chapters 2 and 4. This science is concerned with ground based observations of coherent radio transmissions as they traverse, and thus are dispersed by, the atmosphere and ionosphere of a planet. Information concerning these gases is obtained just before and just after the satellite-borne transmitter-earth-based receiver transmission path is occulted by the planet—that is, near immersion and emersion of the satellite.

The radio-occultation data are inverted to yield profiles of the refractive index, which with appropriate assumptions yield profiles of temperature, pressure, density, and electron density. Inversion algorithms have been developed specifically for this problem; inversion algorithms developed by seismologists are also being used. These methods are described and compared in this chapter and in parts of Chapter 7. The subject of stellar (noncoherent) occultation, which has been used by astronomers to infer the presence and character of planetary and satellite atmospheres, is not specifically discussed in this volume but is obviously related to the present subject.

M. D. Grossi organized and chaired the session devoted to radio-occultation measurements.

CURRENT METHODS OF RADIO OCCULTATION

DATA INVERSION*

Arvydas J. Kliore

Jet Propulsion Laboratory
California Institute of Technology
Pasadena, California

N73-11596

ABSTRACT

The methods of Abel integral transform and ray-tracing inversion have been applied to data received from radio occultation experiments as a means of obtaining refractive index profiles of the ionospheres and atmospheres of Mars and Venus. In the case of Mars, certain simplifications are introduced by the assumption of small refractive bending in the atmosphere. General inversion methods, independent of the thin atmosphere approximation, have been used to invert the data obtained from the radio occultation of Mariner 5 by Venus; similar methods will be used to analyze data obtained from Jupiter with Pioneers F and G, as well as from the other outer planets in the Outer Planet Grand Tour missions.

INTRODUCTION

Since 1965 the method of radio occultation has been used to study the atmospheres of Mars and Venus. The Mariner 4 spacecraft was used to obtain information on the atmosphere of Mars in 1965 [Kliore *et al.*, 1967; Kliore *et al.*, 1968; Fjeldbo and Eshleman, 1968]. In 1967, the Mariner 5 spacecraft flew past Venus and performed both an uplink dual-frequency and a downlink S band occultation experiment [Kliore *et al.*, 1967; Mariner Stanford Group, 1967; Fjeldbo and Eshleman, 1969; Kliore *et al.*, 1969; Fjeldbo *et al.*, 1971]. In 1969, two more Mariner spacecraft flew by Mars and again S band downlink radio occultation experiments were performed [Kliore *et al.*, 1969; 1970 a,b; Fjeldbo *et al.*, 1970]. Similar experiments will be carried out with the Mariner Mars 1971 Orbiter, the Pioneer F & G fly by missions to Jupiter, the Mariner Venus-Mercury fly by, and probably all future Planetary missions, including the Outer Planet Missions.

The analysis and interpretation of the data collected in the radio occultation experiments were carried out on the basis of theoretical foundations established earlier [Fjeldbo, 1964; Fjeldbo and Eshleman, 1965; Fjeldbo *et al.*, 1965; Kliore *et al.*, 1965]. The analysis primarily involved the applications of optical ray-tracing methods under the assumption of spherically symmetrical refractive index distributions. In the case of Mars, an additional simplification accrued from the very small refractive bending produced by a thin atmosphere. The basic methods that are presented here have been used in the past to obtain index of refraction profiles from observed phase and frequency data. In addition, some future applications of these methods are briefly discussed.

RADIO OCCULTATION OBSERVABLES

The quantities observed on earth during a radio occultation experiment are the frequency and amplitude of the microwave radio signal received at a ground station. The amplitude information is important in determining the exact times of loss and reacquisition of the signal, and it can be used to obtain refractive index or absorption profiles. However, the frequency of the signal forms the basic data used to obtain the radial distribution of refractive index. Since the observed frequency includes the effects of spacecraft motion, rotation of the earth, and so forth, it is

*This paper represents the results of one phase of research carried out at the Jet Propulsion Laboratory, California Institute of Technology, under NASA contract NAS7-100.

not immediately suitable for use in analysis. To obtain the frequency change due only to the effects of the planetary atmosphere, the predicted frequency, based on an estimate of the spacecraft orbit, is subtracted from the observed. This is the so-called "frequency residual," which is nothing more than the time rate of change of the total phase change caused by delay and refractive bending in a planetary atmosphere. Thus, in simplest terms, the inversion problem in the reduction of occultation data is simply this: How does one obtain a profile of refractivity as a function of radial distance in the planetary atmosphere from data consisting of the time rate of change of the observed total phase change as a function of time?

To obtain an understanding of the relationship of the observed phase change to the index of refraction, it is of interest to examine first the relationship of the phase path and refractive bending angle to a refractive index profile. Referring to figure 1, the angle of incidence i is related to the ray-path coordinates r and ϕ by:

$$\tan i = r d\phi / dr \quad (1)$$

From Bouger's rule, which applies to spherically stratified index of refraction distributions, one can express the ray asymptote distance a in terms of the index of refraction n and the angle of incidence i :

$$a = nr \sin i \quad (2)$$

From figure 1, it is apparent that

$$\frac{\pi}{2} - \phi = i - \psi \quad d\psi = di + d\phi \quad (3)$$

By differentiating Eq. (1), one may obtain an expression for the differential di as follows:

$$di = \frac{a [n + (dn/dr)r] dr}{nr [(nr)^2 - a^2]^{1/2}} \quad (4)$$

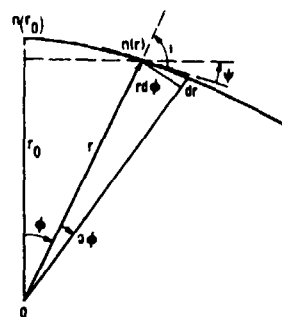


Figure 1.— Ray path geometry.

Similarly by combining Eqs. (1) and (2), one obtains an expression for the differential $d\phi$

$$d\phi = \frac{a dr}{r [(nr)^2 - a^2]^{1/2}} \quad (5)$$

From Eqs. (4) and (5), the expression for $d\psi$ can be obtained as

$$d\psi = \frac{a(dn/dr)dr}{n [(nr)^2 - a^2]^{1/2}} \quad (6)$$

The total bending angle α is then twice the integral of $d\psi$ along the ray path as r goes from r_0 to infinity:

$$\alpha(a) = -2a \int_{r_0}^{\infty} \frac{(dn/dr) dr}{n [(nr)^2 - a^2]^{1/2}} \quad (7)$$

The phase path length ρ along the ray path can be derived as follows. A element of phase path $d\rho$ is given by

$$d\rho = c d\tau \quad (8)$$

where c is the speed of light in free space and $d\tau$ is the differential time of propagation along the ray path element ds . Thus, $d\rho$ can be expressed as:

$$d\rho = c \frac{ds}{v} = n ds \quad (9)$$

With reference to figure 1, it is obvious that

$$ds = [dr^2 + r^2 d\phi^2]^{1/2} \quad (10)$$

Combining Eqs. (10) and (5), one obtains

$$ds = \frac{nr dr}{[(nr)^2 - a^2]^{1/2}}$$

Then, the total phase path length along the ray path, which is equivalent to the apparent distance traveled in free space (taking into account, however, the actual velocity of propagation in the medium, as well as the curvature due to refraction), is given by twice the integral of $d\rho$ from r_0 to infinity:

$$\rho(a) = 2 \int_{r_0}^{\infty} n ds = 2 \int_{r_0}^{\infty} \frac{n^2 r dr}{[(nr)^2 - a^2]^{1/2}} \quad (12)$$

When the frequency residual observed during a radio occultation experiment is integrated from the time of onset of the atmospheric phenomena to the time of interest, the result is the total observed phase change as a function of time:

$$\Phi(t) = \int_{t_0}^t \Delta f(\tau) d\tau \quad (13)$$

The relationship of this phase change to the refraction phenomena in a spherically stratified atmosphere can be explained with the help of the diagram in figure 2. The circular arc of the radius r_m represents the limit of the sensible atmosphere. The value of r_m is arbitrary, the only restriction being that the index of refraction for $r > r_m$ be equal to unity. The phase path change observed at the earth is the difference between the propagation phase path $SA'AE$ and the direct path to the earth $SB'BE$. This phase path change for a one-way propagation through the atmosphere can be represented in terms of the ray asymptote distance a , the refractive bending angle α , and the propagation phase path length ρ as follows:

$$\Phi(a) = \frac{1}{\lambda} [\rho_m(a) + (R^2 - a^2)^{1/2} (1 - \cos \alpha(a)) - 2(r_m^2 - a^2)^{1/2} - a \sin \alpha(a)] \quad (14)$$

where R is the distance from the spacecraft to the center of the planet, ρ_m is

$$\rho_m(a) = 2 \int_{r_0}^{r_m} \frac{n^2 r dr}{[(nr)^2 - a^2]^{1/2}} \quad (15)$$

and λ is the free space wavelength, or the ratio at the speed of light c to the frequency f .

INVERSION PROCEDURE

To convert the phase observable as a function of time into a quantity suitable for inversion, the trajectory, or ephemeris, of the spacecraft must be brought into play. For each time t_i for which the values of Δf and Φ are available, the spacecraft ephemeris provides a position vector relative to the receiving station on the earth, a position vector relative to the center of the planet, and a velocity vector relative to the center of the planet. The component of the planet-centered spacecraft velocity in the direction of the earth, one light propagation interval after the time t_i , is given by

$$\bar{V}_F = \bar{\nabla} \cdot \hat{E} \quad (16)$$

where $\hat{\mathbf{V}}$ is the planet-centered velocity in the plane containing the spacecraft and the centers of the earth and the planet, and $\hat{\mathbf{E}}$ is the unit vector in the direction of the earth one light propagation time after the time t_i . The Doppler frequency that one would expect to see at the earth if propagation were to take place in the direction $\hat{\mathbf{E}}$ is

$$\Delta f_E = \frac{f}{c} |\nabla_E| \quad (17)$$

And the angle ψ_E between the velocity vector \vec{V} and the vector in the direction of the earth, is given by (see figure 2)

$$\psi_E = \cos^{-1} \frac{|\vec{V}_E|}{|\vec{V}|} \quad (18)$$

However, the angle measured between the velocity vector of the spacecraft and the actual direction of the ray that ultimately arrives at the earth after undergoing refraction in the planetary atmosphere, is given by [Phinney and Anderson, 1968].

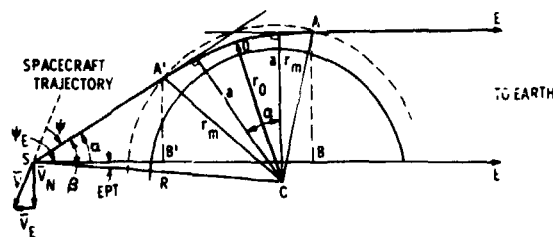


Figure 2.— Geometry for the computation of phase change.

$$\psi = \cos^{-1} \left[\frac{c}{\lambda |\nabla|} (\Delta f_E + \Delta f) \right] \quad (19)$$

where Δf is the observed frequency residual. The refractive bending angle α is then given by

$$\alpha = \psi_F^\dagger \psi \quad (20)$$

At the same time, it can be seen from figure 2 that the ray asymptote distance a can be obtained by

$$a = R \sin \beta \quad (21)$$

where $\beta = \alpha + EPT$, and EPT = the angle subtended by the earth and the center of the planet at the spacecraft.

Now one has both the value of the refractive bending angle α and the ray asymptote distance a for every input data point – or in effect, the refractive bending angle as a function of the ray asymptote distance. Furthermore, one can now relate the observed total phase change Φ to the appropriate value of the ray asymptote distance a , and

hence one has Φ as a function of a . It is a very simple matter to also obtain the expression for the phase path change in the atmosphere ρ_m . From Eq. (14), it follows that

$$\rho_m(a) = \lambda \Phi(a) - (R^2 - a^2)^{1/2} [1 - \cos \alpha(a)] + 2(r_m^2 - a^2)^{1/2} + a \sin \alpha(a) \quad (22)$$

Thus, one now has computed the values of the refractive bending angle α and the phase path ρ as functions of the ray asymptote distance a .

Thin Atmosphere Approximation

In a thin atmosphere, such as that of Mars, the refractive bending angle is very small, and in that case Eq. (19) can be written as follows:

$$\cos \psi = \cos \psi_E + \frac{\lambda}{|\bar{V}|} \Delta f \quad (23)$$

But since $\psi = \psi_E + \alpha$, and α is very small

$$\alpha = -\frac{\lambda}{|\bar{V}_N|} \Delta f \quad (24)$$

where \bar{V}_N is the velocity of the spacecraft in the plane of the spacecraft, earth and planet centers, and normal to the direction of the earth. Similarly, the equation for the ray asymptote distance a becomes

$$a = R(\sin EPT + \alpha \cos EPT) \quad (25)$$

The expression for the total phase change can then be written as

$$\Phi(a) = \frac{1}{\lambda} \left\{ \left[\rho_m(a) - 2(r_m^2 - a^2)^{1/2} - a \alpha(a) \right] + (R^2 - a^2)^{1/2} \frac{\alpha^2(a)}{2} \right\} \quad (26)$$

Now note that the path of integration A' to A in figure 2 becomes a straight line for a thin atmosphere, such that

$$AA' = AO + OA' = 2(r_m^2 - a^2)^{1/2} + a\alpha = 2 \int_{r_0}^{r_m} ds \quad (27)$$

Hence, Eq. (26) can be expressed as

$$\Phi(a) = \frac{1}{\lambda} \left[\rho_1(a) + (R^2 - a^2)^{1/2} \frac{\alpha^2(a)}{2} \right] \quad (28)$$

where

$$\begin{aligned}\rho_1(a) &= 2 \int_{r_0}^{r_m} \frac{(n-1)nr \, dr}{[(nr)^2 - a^2]^{1/2}} \\ &= \lambda \Phi(a) - (R^2 - a^2)^{1/2} \frac{\alpha^2(a)}{2}\end{aligned}\quad (29)$$

Also, since in a thin atmosphere the refractive index n is very nearly equal to 1, the quantity ρ_1 can be expressed as

$$\begin{aligned}\rho_1(a) &\simeq 2 \times 10^{-6} \int_{r_0}^{r_m} \frac{N(r)r \, dr}{(r^2 - a^2)^{1/2}} \\ N &= (n-1) \times 10^6\end{aligned}\quad (30)$$

Now, introducing the transformation of variables

$$\begin{aligned}y &= r_m^2 - r^2 \\ w &= r_m^2 - a^2\end{aligned}\quad (31)$$

The integral in Eq. (30) can be expressed as

$$\rho_1(w) = 10^{-6} \int_0^w \frac{N(y) \, dy}{(w-y)^{1/2}} \quad (32)$$

This expression is now in a form suitable to inversion by means of the Abel integral transform [Hamel, 1937], which states that if

$$f(w) = K \int_0^w \frac{g'(y) \, dy}{(w-y)^{1/2}} \quad (33)$$

then

$$g(y) = \frac{1}{k\pi} \int_0^y \frac{f(w) \, dw}{(y-w)^{1/2}} + g_0 \quad (34)$$

Applying the transform to Eq. (32), one obtains:

$$N(y) = \frac{10^6}{\pi} \frac{d}{dy} \int_0^y \frac{\rho_1(w) \, dw}{(y-w)^{1/2}} \quad (35)$$

and

$$rN(r) = \frac{2 \times 10^6}{\pi} \frac{d}{dr} \int_{r_m}^r \frac{a \rho_1(a) da}{(a^2 - r^2)^{1/2}} \quad (36)$$

Now performing the change of variables

$$a = r \cosh \theta \quad (37)$$

and performing the differentiation with respect to r , the following result is finally obtained:

$$N(r) = \frac{10^6}{\pi r^2} \int_{r_m}^r \frac{a [\rho_1(a) + a(d\rho_1(a)/da)] da}{(a^2 - r^2)^{1/2}} \quad (38)$$

The expression in Eq. (38) was used to produce the refractivity profiles from the radio occultation data obtained with the Mariner 6 and Mariner 7 spacecraft in 1969. The original data, as well as some of the intermediate products of the Mariner 6 data reduction, are shown in figures 3, 4, and 5; the resulting refractivity profile in figure 6. The residual frequency change Δf for a two-way passage through the atmosphere and ionosphere of Mars is shown in figure 3. The total phase change Φ is shown as a function of time in figure 4. Figure 5 shows the corrected phase path change in the atmosphere of Mars (in km) for a one-way passage as the function of the ray asymptote distance a . The same basic inversion method will be used to analyze the data obtained from the Mariner Mars 1971 orbiter.

An inversion procedure making use of the thin atmosphere approximation and an assumption of a spherically stratified atmosphere divided into a finite number K of concentric layers was used by *Fjeldbo and Eshleman* [1968] to invert the occultation data obtained from the Mariner 4 spacecraft in 1965. Referring to figure 7, Eq. (30) can be expressed as follows

$$10^6 \rho_1(m) = N(m) \Delta z(m, n) + 2 \sum_{n=1}^{m-1} N(n) \Delta z(m, n) \quad (39)$$

Where $\rho_1(m)$ denotes the phase path change for the m th ray, and $N(m)$ is the refractivity in the m th layer. The element $\Delta z(m, n)$ is the length of the portion of the m th ray that lies in the n th layer (figure 7). The two-dimensional element Δz is related to the radii of the atmospheric layers by

$$\begin{aligned} \Delta z(m, n) &= \left[r_n^2 - \left(\frac{r_m + r_{m+1}}{2} \right)^2 \right]^{1/2} - \left[r_{n-1}^2 - \left(\frac{r_m + r_{m+1}}{2} \right)^2 \right]^{1/2} \\ \Delta z(m, m) &= 2 \left[r_m^2 - \left(\frac{r_m + r_{m+1}}{2} \right)^2 \right]^{1/2} \end{aligned} \quad (40)$$

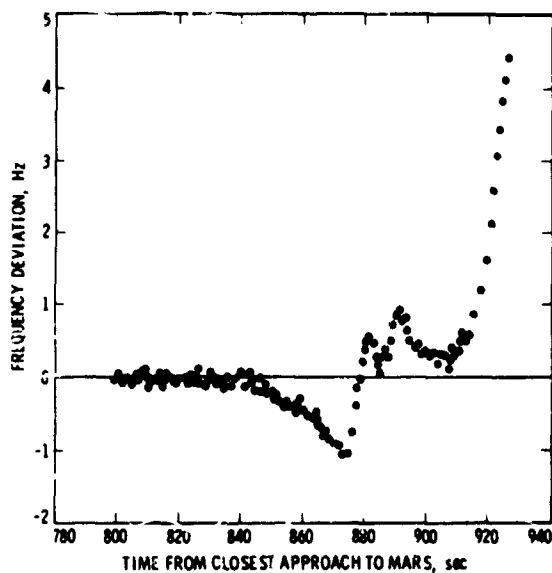


Figure 3.— Frequency residuals Δf from Mariner 6 entry data.

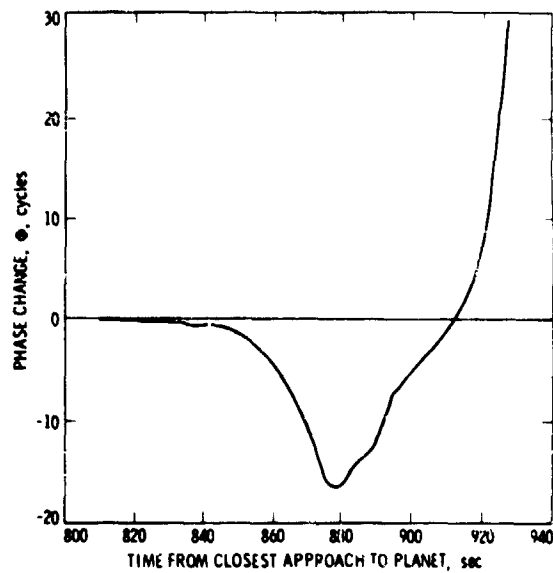


Figure 4.— Total phase change, Mariner 6 entry.

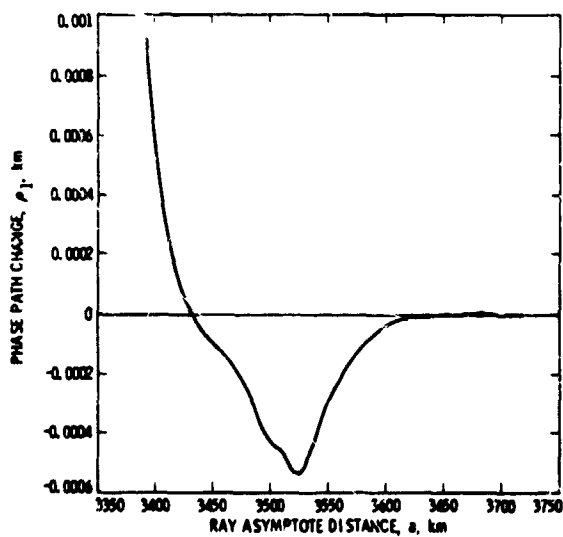


Figure 5.— Phase path change in the atmosphere, Mariner 6 entry.

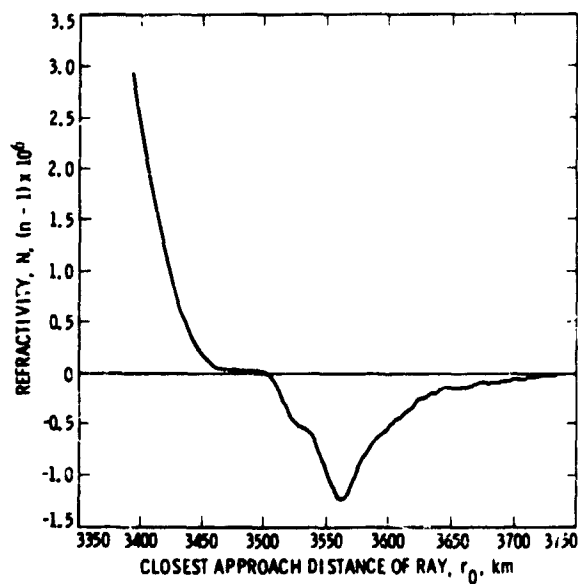


Figure 6.— Refractivity profile obtained by integral inversion, Mariner 6 entry.

Eq. (39) can be repetitively solved for N , yielding the following procedure:

$$\begin{aligned}
 N(1) &= \frac{10^6 \rho_1(1)}{\Delta z_{11}} \\
 N(2) &= \frac{10^6 \rho_1(2) - 2N(1)\Delta z(2,1)}{\Delta z(2,2)} \\
 N(m) &= \frac{10^6 \rho_1(m) - 2 \sum_{n=1}^{m-1} N(n)\Delta z(m,n)}{\Delta z(m,m)} \quad (41)
 \end{aligned}$$

The procedure (41) provides the refractivity of the m th layer as a function of the radius of the layer; figure 8 shows the refractivity profiles obtained from Mariner 4 radio occultation measurement [Fjeldbo and Eshleman, 1968] by means of this method.

The simplified inversion procedures making use of the thin atmosphere approximations are applicable to the analysis of data obtained from Mars. However, for planets such as Venus and Jupiter, which have dense atmospheres, and in which refractive bending is quite considerable (approximately 17° near the super-refractive level on Venus), more general methods of inversion must be used.

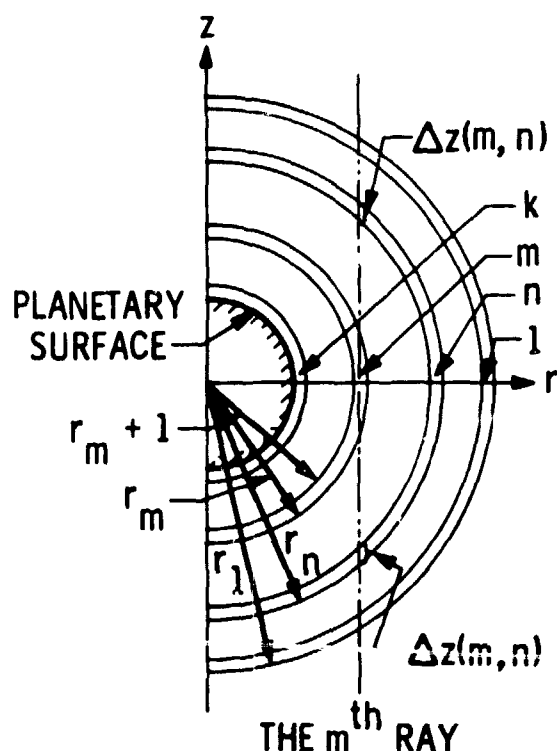


Figure 7. - Cross section of atmosphere consisting of K spherical layers [From: Fjeldbo and Eshleman, 1968].

General Methods

It has been shown how the refractive bending angle α and the phase path change ρ_1 can be obtained as functions of the ray asymptote distance a from the observed frequency residual Δf and the ephemeris of the spacecraft. Both α and ρ_1 , as well as a number of other related quantities, can be inverted to obtain the refractive index profile by means of the Abel integral transform. Let us first consider the inversion of the refractive bending angle α .

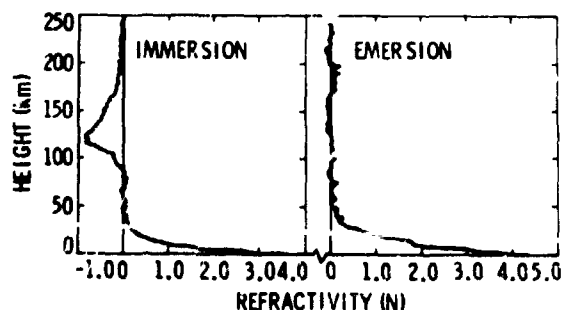


Figure 8. - Mariner 4 refractivity profiles [From: Fjeldbo and Eshleman, 1968].

In a method described by *Fjeldbo et al.* [1971], Eq. (7) is first transformed by replacing the product nr by a variable of integration x :

$$\alpha(a) = -2a \int_a^\infty \frac{dn}{n dx} \frac{dx}{(x^2 - a^2)^{1/2}} \quad (42)$$

Now letting a_1 denote the ray asymptote distance for a closest approach distance r_{O1} , we multiply both sides of Eq. (42) by the quantity $(a^2 - a_1^2)^{1/2}$ and integrate with respect to a from a_1 to infinity. This yields:

$$\int_{a_1}^\infty \frac{\alpha(a) da}{(a^2 - a_1^2)^{1/2}} = \pi \ln n(r_{O1}) \quad (43)$$

The term on the left-hand side can be integrated by parts, resulting in

$$n(r_{O1}) = \exp \left[\frac{1}{\pi} \int_0^{\alpha(a_1)} \ln \left(\frac{a(\alpha)}{a_1} + \left\{ \left[\frac{a(\alpha)}{a_1} \right]^2 - 1 \right\}^{1/2} \right) d\alpha \right] \quad (44)$$

which is an expression for the refractive index at the closest approach radius of the ray r_{O1} given by

$$r_{O1} = \frac{a_1}{n(r_{O1})} \quad (45)$$

The phase path change ρ_m can be inverted in a similar manner [*Phinney and Anderson*, 1968]. Again letting $x = nr$, Eq. (15) becomes:

$$\rho_m(a) = 2 \int_a^{r_m} \frac{x^2}{r(x^2 - a^2)^{1/2}} \frac{dr}{dx} dx \quad (46)$$

Now, performing the substitution of variables:

$$y = r_m^2 - x^2 \quad w = r_m^2 - a^2 \quad (47)$$

One obtains the following expression for ρ .

$$\rho_m(w) = -2 \int_0^w \frac{(r_m^2 - y) [d(\ln r)/dy] dy}{(w - y)^{1/2}} \quad (48)$$

Again making use of the Abel integral transform [Eqs. (33) and (34)] one has:

$$g'(y) = (r_m^2 - y) \frac{d(\ln r)}{dy} = \frac{d}{dy} \left[-\frac{1}{2\pi} \int_0^y \frac{f(w)}{(y-w)^{1/2}} dw \right] \quad (49)$$

or, reverting to the original variables x and a ,

$$\frac{x^2 d(\ln r)}{dx} = \frac{d}{dx} \left[\frac{1}{\pi} \int_{r_m}^x \frac{a \rho_m(a) da}{(a^2 - x^2)^{1/2}} \right] \quad (50)$$

It can be shown [Phinney and Anderson, 1968] that this expression ultimately yields

$$r(x) = r_m \exp \left[-\frac{1}{\pi} \int_{r_m}^x \cosh^{-1} \frac{a}{x} \frac{1}{a} \frac{d\rho_m}{da} da \right] \quad (51)$$

and finally, the index of refraction as a function of r (the closest approach distance of the ray) can be obtained by

$$n(r) = \frac{x}{r(x)} \quad (52)$$

We have seen that both the refractive bending angle α and the phase path change ρ_1 can be inverted to yield the refractive index n as a function of the closest approach distance of the ray r by means of the Abel integral transform, under the assumption of a spherically symmetrical index of refraction profile.

Ray-tracing inversion may also be used to obtain the index of refraction profile from data consisting of the refractive bending angle as a function of ray asymptote distance [Fjeldbo, et al., 1971].

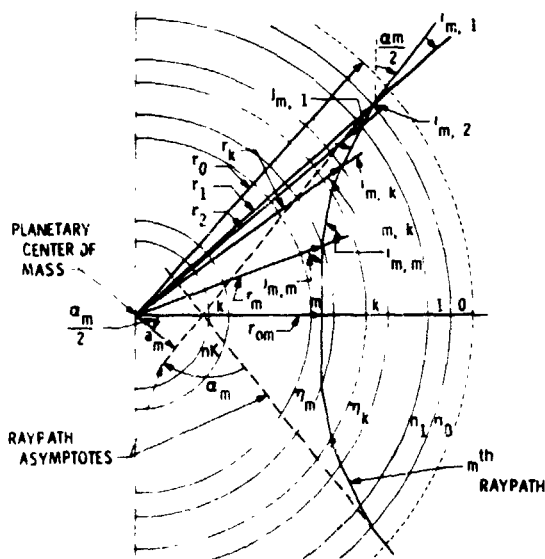


Figure 9.— Refraction of ray path by concentric spherical layers — index of refraction increasing with height [From: Fjeldbo, Kliore, and Eshleman, 1971].

Figure 9 shows the geometry for a ray that is being refracted in an ionosphere consisting of K concentric layers. If r_m and n_m are the radius and the index of refraction respectively, for the m th layer, then the radius of closest approach r_{om} of the m th ray is given by

$$r_{om} = \frac{1}{2} (r_m + r_{m+1}) \quad (53)$$

If $i_{m,k}$ is the angle of incidence and, $j_{m,k}$ is the angle of refraction at each boundary, then the law of sines and Snell's law applied at the first boundary gives the following equations for the m th ray:

$$a_m = r_1 \sin i_{m,1} \quad (54)$$

$$\sin i_{m,1} = n_1 \sin j_{m,1} \quad (55)$$

$$\psi_m = \frac{1}{2} \alpha_m + i_{m,1} - j_{m,1} \quad (56)$$

where a_m = ray asymptote distance for the m th ray

α_m = total angle through which the m th ray is bent by the atmosphere

$\psi_{m,1}$ = the angle which the m th ray must be bent before it is traced from the first layer to the closest approach point.

Generalizing to the k th boundary gives

$$r_{k-1} \sin j_{m,k-1} = r_k \sin i_{m,k} \quad (57)$$

$$n_{k-1} \sin i_{m,k} = n_k \sin j_{m,k} \quad (58)$$

$$\psi_{m,k} = \psi_{m,k-1} + i_{m,k} - j_{m,k} \quad (59)$$

The data giving α as a function of a , and the equations given above may be used to determine the refractive index n as a function of r for each layer. For a ray passing above the atmosphere, the bending angle is zero and the radius of closest approach is equal to the asymptote distance. For this layer, then,

$$n_0 = 1; \quad r_0 = \frac{1}{2}(a_0 + a_{-1}) \quad (60)$$

where a_0 and a_{-1} denote the asymptote distances for the rays that pass immediately above the first atmospheric layer. Then from Eq. (51) the radius of the first layer is:

$$r_1 = 2a_0 - r_0 \quad (61)$$

Setting m equal to 1 and $\psi_{1,0}$ equal to $\alpha_1/2$ in the ray equations, we can calculate the refractive index for the first layer:

$$n_1 = a_1 \left[r_1 \sin \left(\frac{\alpha_1}{2} + \sin^{-1} \frac{a_1}{r_1} \right) \right] \quad (62)$$

The radius of closest approach for the first ray r_{01} is given by

$$r_{0m} = r_m \sin j_{m,m}, \quad m = 1 \quad (63)$$

Again applying Eq. (53) with $m = 1$, we can determine the radius of the second layer r_2 . Continuing these computations for the remaining rays, the entire refractive index profile of the atmosphere is obtained. For instance, the radius r_m and the refractive index n_m of the m th layer are determined from the $(m-1)$ st and the m th ray, respectively. The general expressions are of the form

$$r_m = 2r_{0m} - r_{m-1} \quad (64)$$

$$n_m = n_{m-1} \frac{r_{m-1} \sin i_{m,m-1}}{r_m \sin \left\{ \psi_{m,m-1} + \sin^{-1} \left[(r_{m-1}/r_m) \sin i_{m,m-1} \right] \right\}} \quad (65)$$

With the introduction of additional mathematical complexity, the ray-tracing inversion method can be generalized to include cases in which the refractive index is allowed to change in both vertical and horizontal directions. However, when only a single occultation is observed, the horizontal changes must be specified before the data can be utilized to determine the vertical refractive index profile.

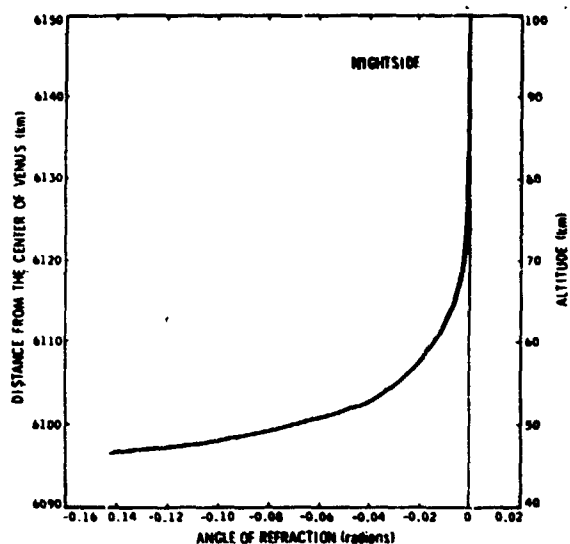


Figure 10. - Refractive bending angle versus the altitude at the ray asymptote for Mariner 5 entry, computed from 423.3 MHz amplitude data [From: *Fjeldbo, Kliore and Eshleman, 1971*].

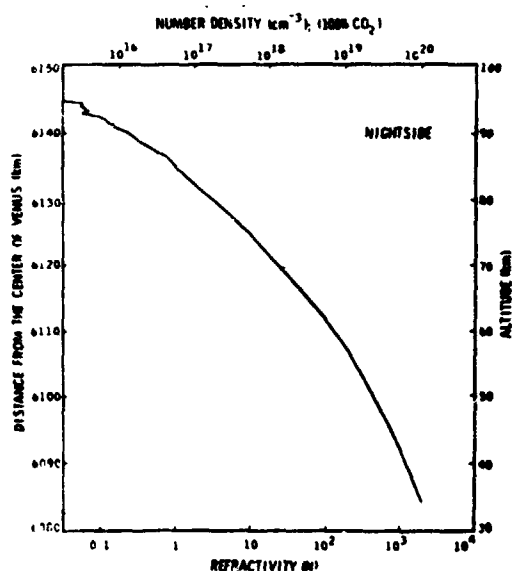


Figure 11. - Refractivity and number density profile for Mariner 5 entry [From: *Fjeldbo, Kliore and Eshleman, 1971*].

Fjeldbo has applied the ray-tracing inversion method to analyze the data from the occultation of Mariner 5 by Venus [Fjeldbo, et al., 1971]. The refractivity profile resulting from the inversion of the Mariner 5 Venus night-side data is shown in figure 11.

Mathematical techniques similar to those outlined for the inversion of radio occultation data from planets having dense atmospheres, already applied to Venus occultation data, will be used in the future to analyze the data obtained at Venus during the Venus-Mercury 1973 mission, and from Jupiter during the Pioneer F and G missions, as well as the Outer Planet Grand Tour Missions which are planned for the latter part of this decade and the 1980's.

AMPLITUDE DATA

Amplitude data collected during a planetary occultation also contain other valuable information. If the frequency is low enough that absorption effects can be neglected, the amplitude data can be assumed to contain information only about refractive defocusing and a procedure exists to deduce the ray bending angle α and the ray asymptote parameter a . Fjeldbo et al., [1971] have shown that starting with rays that graze the top of the atmosphere, the amplitude measurements can be used point by point to solve for the differential change in the angle of refraction $\Delta\alpha$ with decreasing ray asymptote altitude. Integrating $\Delta\alpha$ yields α as a function of a . Figure 10 shows an example of a refractive bending angle profile obtained in this manner from the Mariner 5 Venus 423.3 MHz amplitude data.

On the other hand, when measurements are made at a frequency at which absorption effects in a planetary atmosphere are apparent, the amplitude data can be used in conjunction with the results of the previous section to compute a profile of the absorption coefficient in the atmosphere.

It can be shown from the geometry of figure 2 that the attenuation due to defocusing in the planetary atmosphere (in dB) is given by

$$A_D(a) = 10 \log_{10} A_N(a) \quad (66)$$

where

$$A_N(a) = 1 - \left[(R^2 - a^2)^{1/2} + a \tan \frac{\alpha(a)}{2} \right] \frac{d\alpha(a)}{da} \quad (67)$$

The refractive bending angle α as a function of a is available from the techniques described in the first section. Since the observed total attenuation A_T is available as a function of time, it can be expressed as a function of the ray asymptote distance a by comparison of the computed value of a to the time of receipt of the frequency data. Thus the excess attenuation, or loss A_E , is simply the difference between the total observed attenuation A_T and the computed defocusing attenuation A_N . Now, the excess attenuation due to absorption along the ray path can be described as follows:

$$A_E(a) = (10 \log_{10} e) \tau(a) \quad (68)$$

or

$$\tau(a) = \frac{1}{10 \log_{10} e} A_E = 2 \int_{r_0}^{r_m} \sigma(r) ds \quad (69)$$

where $\sigma(r)$ is the absorption coefficient.

By means of ray-tracing inversion discussed in the previous section, the quantity τ for the m th ray having a closest approach radius of r_{0m} (fig. 9) can be expressed as

$$\tau_m = 2\Delta z_{m,m} \sigma_m + 2 \sum_{k=0}^{m-1} \Delta z_{m,k} \sigma_k \quad (70)$$

where

$$\Delta z_{m,m} = (r_m^2 - r_{0m}^2)^{1/2} \quad (71)$$

and

$$\Delta z_{m,k} = \frac{r_{k+1} \sin(i_{m,k+1} - i_{m,k})}{\sin i_{m,k}} \quad (72)$$

Thus, starting with layer o in which $n_o = 1$

$$\tau_o = 2\Delta z_{o,o} \sigma_o ; \quad \sigma_o = \frac{\tau_o}{2\Delta z_{o,o}} ; \quad \Delta z_{o,o} = (r_o^2 - r_{0,o}^2)^{1/2} \quad (73)$$

In the first layer

$$\tau_1 = 2\Delta z_{1,1}\sigma_1 + 2\Delta z_{1,0}\sigma_0$$

$$\sigma_1 = \frac{\tau_1 - 2\Delta z_{1,0}\sigma_0}{2\Delta z_{1,1}} \quad (74)$$

Finally, generalizing to the m th ray,

$$\sigma_m = \frac{\tau_m - 2 \sum_{k=0}^{m-1} \Delta z_{m,k} \sigma_k}{2\Delta z_{m,m}} \quad (75)$$

The quantities r_{ok} , $i_{m,k}$ and $i_{m,k+1}$ are computed at each step from Eqs. (57), (58), (59), (63), and (64).

A similar method has been used by *Fjeldbo, et al.* [1971] for computing the absorption coefficient in the lower atmosphere of Venus using the Mariner 5 S band attenuation data. It is expected that an application of similar methods to the S band data derived from the Pioneer F and G Jupiter missions will yield profiles of ammonia abundance in those portions of the Jovian atmosphere probed by the Pioneer radio occultation experiment.

DISCUSSION

Parker: If there is significant absorption in the atmosphere, the dispersion relations will imply phase distortions. Doesn't this affect your refractive index profile?

Kliore: Yes. It would actually affect the temperature and pressure profiles.

Parker: If there is significant absorption, the phase distortion is due to two effects: pure refraction and a dispersion relationship.

Kliore: That is true. I don't know the magnitude of the effect, but it will show up in the profile and the profile will be difficult to interpret.

Bojarski: If you have absorption and phase data shouldn't you combine them and pose the inverse problem for a complex index of refraction?

Kliore: That would be difficult since amplitude data include effects of defocusing and absorption.

Reinisch: When you calculated the phase path change you took the difference between the refracted and vacuum paths forming the difference under the integral sign. You should take the difference between two integrals.

Kliore: There is only a notation mistake. We do what you suggest.

Unidentified Speaker: You have described two inversion methods — the Abel transform and the Shell method. Have you compared these?

Kliore: They are quite similar. Each data point is treated as the origin of every shell.

Unidentified Speaker: What is relative importance of defocusing and absorption?

Kliore: Venus is our only source of data. Defocusing predominates. There is no appreciable absorption down to where the defocusing distortion is about 30 dB. This continues until the sum of the two is about 45 dB. The maximum amount of absorption is about 6 dB.

Whitten: We looked at the effect of ionospheric absorption. At your frequencies, the effect is negligible.

Kliore: We have not specifically looked at that. Perhaps it would be important at Jupiter.

Rodgers: Do you have a well-conditioned linear problem, and can you get out as many independent points as you measure?

Kliore: Yes. The problem is solving a first-order linear differential equation. The discrete method is essentially a matrix inversion. The noise problem is manageable.

Unidentified Speaker: In fact, can it be written just as the inversion of an upper triangular matrix?

Kliore: Yes.

**ERRORS INCURRED IN PROFILE RECONSTRUCTION AND
METHODS FOR INCREASING INVERSION ACCURACIES FOR
OCCULTATION TYPE MEASUREMENTS**

S. H. Gross and J. A. Pirraglia

Polytechnic Institute of Brooklyn

N73-11597 ABSTRACT

The atmosphere and ionosphere of a planet varies with latitude, longitude, altitude, and time. The occultation experiment scans the medium roughly along the locus of nearest points of the ray. The present method of measuring and inverting the data is satisfactory if this locus is along a vertical, a condition that is closely approximated when the motion of the spacecraft about the planet is in or close to the ecliptic. When inclined to the ecliptic (as for example, an orbiter, to take advantage of progression or retrogression of the orbit), the resulting profiles are considerably in error with respect to the vertical profiles over the occultation points. Even along the locus of nearest points, significant errors occur about and below the peak of the ionosphere. The resulting profile along the slant locus is difficult to interpret in terms of the physical theories of atmospheres or ionospheres. Proper interpretation requires the components of the gradient with radius and angles.

A method for augmenting the occultation experiment is described for slightly refractive media. This method, which permits separation of the components of the gradient of refractivity, appears applicable to most of the planets for a major portion of their atmospheres and ionospheres. The analytic theory is given, and the results of numerical tests with a radially and angularly varying model of an ionosphere are discussed.

INTRODUCTION

As Mariners 4 through 7 moved behind the planets Venus and Mars, they provided information on the vertical distribution of ionization and neutral gases above points of occultation [Fjeldbo and Eshleman, 1968; Kliore *et al.*, 1965, 1967, 1969]. The vehicles moved in planes close to the ecliptic, and as shown by Pirraglia and Gross [1970], radial profiles should be expected for such paths.

Actual atmospheres and ionospheres vary with latitude and longitude, the time of the planet's day, and altitude or radius. Therefore, more complete comprehension of the physics and properties of the media requires information on the variations with all positional coordinates and time. If attempts are made to obtain such data from spacecrafts in orbits inclined to the ecliptic, using progression or retrogression of the orbital planes to advance or retard the points of occultation, then, as shown by Pirraglia and Gross [1970], the presently constituted experiment yields results considerably in error if interpreted as vertical profiles over the occultation points. At best, the results represent a somewhat inaccurate measurement of the refractivity profile along a slanted locus of nearest points to the planet of rays between the observation point on Earth and the spacecraft. The greater the inclination of the orbit and the angular variation of the medium, the greater is the error. The effects of angular dependence enter because the radiation is scanning the medium in angle as well as radius as a result of the spacecraft's inclined trajectory.

Even if refractivity information along a slant path is considered adequate and sufficiently accurate, it is presently of limited use because most theoretical work on atmospheres and ionospheres has been concentrated on the nature and interpretation of vertical profiles. Adequate theory would be needed for comprehending and interpreting combined vertical and horizontal variations. Difficulties could be encountered with special situations, such as when the slant locus

lies tangent to surfaces of constant refractivity. Since the medium is not stationary in time, these problems are not alleviated by the accumulation of data from numerous occultations of an orbiter in slightly different regions.

The effect of angular variation of refractivity on the occultation experiment depends on the ratio of the planet's radius to the scale height. The larger this ratio, the less important is the contribution of latitudinal and longitudinal gradients on the measurement. This is because the angular scan of the nearest point locus is smaller. Since the lower neutral atmosphere usually has a scale height that is much smaller than ionospheric scale heights, it is expected that difficulties will be greater for ionospheres. If angular gradients in the neutral atmosphere are greater in the ratio of the scale heights of the two regions of the medium surrounding a planet, then the difficulties will be comparable. Deep neutral atmospheres, such as for Venus and very likely for Jupiter and Saturn, produce large refractive effects that cause considerable bending of the rays, and may cause skewing and torsion. The locus of nearest points may be considerably distorted and extend the angular scan of the medium.

It is desirable to measure the radial and angular components of the gradients of refractivity. These components, measured for each point along the locus of nearest points, would at least permit some reconstruction of the profile along radial or vertical lines at various places along the locus. A complete reconstruction would require the determination of higher order derivatives as well. Although in principle these derivatives are measurable, they would be exceedingly difficult to determine in practice. Thus, only first-order partial derivatives may be sought, and reconstruction is reasonable only to the extent that expansion to first-order terms is acceptable. In this sense, the medium would be best if it were slowly varying with angular coordinates. Rapid changes with angle of heights of separate distinct layers would cause difficulties. At worst, knowledge of the components of the gradients would be useful by itself for more complete theoretical treatment of the likely physical processes controlling the regions.

Here we present a method for measuring both radial and angular gradients by means of a modified occultation experiment. The method utilizes simultaneous scans of the medium along slightly different loci. It has only been applied to slightly refractive media. Some of the ideas may be applicable to highly refractive regions, but such cases require further study. One must utilize dispersion for scanning on slightly different loci. For an ionosphere, frequency dispersion is readily available. For neutral media, such dispersion is only possible at millimetric waves where the experiment would suffer from absorption. To avoid severely attenuated signals much lower frequencies must be used where frequency dispersive effects are negligible. Dispersion must then be spatial, as may be possible with closely spaced radiators emitting at different frequencies to avoid interference effects. It should be noted that an experiment has been suggested for performing occultations between two spacecraft rather than to earth, which has some merits in view of the difficulties described here [Grossi *et al.*, 1969].

The analysis presents the theory of the method of measurement as applied to an ionosphere model simulating a medium with both angular and radial variation. Exact values for the refractivity are known and are compared with computed values deduced from simulated measurements. The results of these calculations are presented. The angular derivatives transverse to the rays are most easily obtained. Angular derivatives along the ray are much more difficult to measure. If the locus of nearest points, from one end to the other, does not cover too large a change in the angular coordinate parallel to the rays, the determination of these derivatives is not essential. For media that do not extend too far from the planet (relative to the radius of the planet), the major angular change appears to be transverse to the rays, a situation that may be characteristic for real ionospheres.

THEORY

The residual Doppler shift of a signal passing through a medium from a moving spacecraft depends on all components of the gradient of refractivity of the medium. If the contributions of the radial and transverse components are separable, it is possible to obtain approximations to these components of refractivity by means of Abel transformations. The refractivity and its derivatives, so obtained, are properties of the medium at the nearest points.

When this locus is close to a vertical to the planet's surface, the profile obtained by direct inversion of the residual Doppler data is a good approximation to the vertical profile, with the possible exception of complications below an ionization peak, or for regions of large gradients. Such direct inversion is the procedure based on spherical symmetry that has been applied to the Mariner experiments [Fjeldbo, 1964; Fjeldbo *et al.*, 1965; Fjeldbo and Eshleman, 1965].

When the locus deviates appreciably from a vertical line, the assumption that the derived data is a vertical profile is incorrect, as shown by Pirraglia and Gross [1970], for media with transverse gradient components and even for media with large radial components of the refractivity gradient.

In the occultation experiment one may view the locus of nearest points as a line scanned by the experiment, the scanning being accomplished by the motion of the spacecraft along its trajectory. None of the components of the gradient of refractivity may be separated from one such line, unless it is a coordinate line, such as a radial or a circle of constant radius. The only derivable information is an approximate directional derivative along the locus.

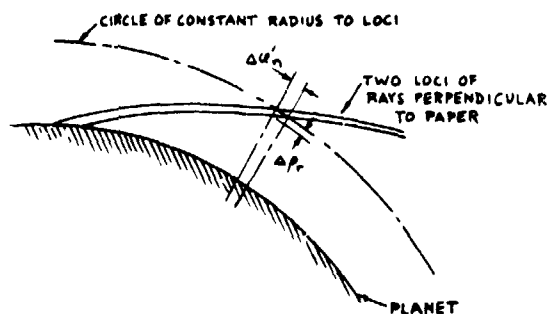


Figure 1.— Nearest point loci. The loci of nearest points of two sets of rays are shown projected into the plane perpendicular to the direction of observation.

One may appreciate in a qualitative sense that it is possible to obtain information on the components of the gradient by simultaneously utilizing more than one scanning line with small separations between such lines. These lines as viewed from the observation point appear to be projected into a plane perpendicular to the line of observation (Fig. 1). The spacing between corresponding points on these lines can be related to the radial component of the refractivity gradient and an angular component of the gradient, the component transverse to the radial component as seen in the projected plane. In terms of a cylindrical coordinate system with the polar axis along the direction of observation, the z_r axis, and the coordinates of the plane of projection designated ρ_r and ϕ_r , these components of the refractivity would be $\partial N / \partial \rho_r$ and $1 / \rho_r \partial N / \partial \phi_r$. The component $\partial N / \partial z_r$, which is essentially parallel to the rays, appears to be poorly defined by the loci.

This pictorial explanation is not exact, and it turns out that one may obtain the component $\partial N / \partial z_r$, but the sensitivity required to accomplish this is far greater than for the other components. In principle, the more loci scanned the greater the number and order of derivatives that may be obtained. For slightly refractive media it turns out that z_r over the full sweep of the loci barely changes, so that the principal changes occur in the coordinates ρ_r and ϕ_r .

The component $\partial N / \partial \phi_r$ is of great value in reconstructing a profile along a vertical, as well as in providing information on the nature of the change of the medium with angle. It is of interest that if the loci lie tangent to contours of constant refractivity, determination of $\partial N / \partial \rho_r$ and $\partial N / \partial \phi_r$ will clarify this situation.

The derivation of analytic expressions requires the assumption that the medium has small refractivity and that the index of refraction μ may be expressed as

$$\mu = 1 + \epsilon N \quad (1)$$

where ϵN is the refractivity ($\ll 1$). ϵ is a small constant parameter, and N is the part of the refractivity that contains the dependence on position coordinates r, θ, ϕ . N may be of the order of unity in regions of maximum medium density but may decrease exponentially as density decreases.

It is also assumed that the variation of N with θ and ϕ is small enough that N may be expanded in a Taylor series about some coordinate angles θ_0, ϕ_0 , and in a restricted region about these values the expansion may be cut off at the first-order derivative terms.

Though these assumptions are needed to permit the derivation of analytic expressions, the method is applicable to a wider range of parameters when treated by computational techniques. The Taylor series expansion need not be cut off at the first-order derivatives. Higher order derivatives may be analytically treated by utilizing additional measurements, but the practicality of such measurements could present severe equipment problems.

If $R(t)$ is the radius vector of the spacecraft's path, measured from some convenient origin, such as the planet's center, the residual phase length of the rays outside the region of the caustic has a one-to-one association with

the vector function \mathbf{R} and may be expressed as $\Delta\Phi[\mathbf{R}(t)]$. For these conditions there is also a single-valued relationship between points on the locus of nearest points $r_n(t)$, and the path of the spacecraft, so that $\Delta\Phi = \Delta\Phi[r_n(t)]$. The residual phase length may then be expressed either in terms of the coordinates of the spacecraft's position or in terms of the coordinates of the nearest points of the rays.

The position of the spacecraft will be specified in terms of the cylindrical coordinate system, that has been defined. The spacecraft's coordinates are designated ρ_s, ϕ_s, z_s ; ϕ_s is measured from some convenient x_r axis such as the ecliptic north direction, assuming that z_r axis lies in the ecliptic plane. (One may readily modify these definitions when a planet's position out of the ecliptic plane is taken into account.) This coordinate system is considered fixed for the occultation. (Changes in the earth's position are not important for the presentation here.) Another coordinate system may be defined for the nearest point of the ray emanating from or being received from the spacecraft in any one position along its trajectory. This coordinate system is rotated relative to the fixed system. Its polar or z-axis, designated z_n , is parallel to the tangent to the ray at the nearest point and is in the same general direction as the direction of observation. Its x-axis, designated x_n , is the radial from the center of the planet, the origin of this coordinate system, through the nearest point. The radius to the nearest point is designated r_n , and this axis is located relative to the fixed axes by an azimuthal plane rotation ϕ'_n about the z_r axis and by a rotation θ'_n in this plane. The coordinate systems and their relationship are shown in figure 2. The coordinates r_n, θ'_n, ϕ'_n for the nearest point is actually spherical coordinates for the nearest point with θ'_n the complement of the usual polar angle.

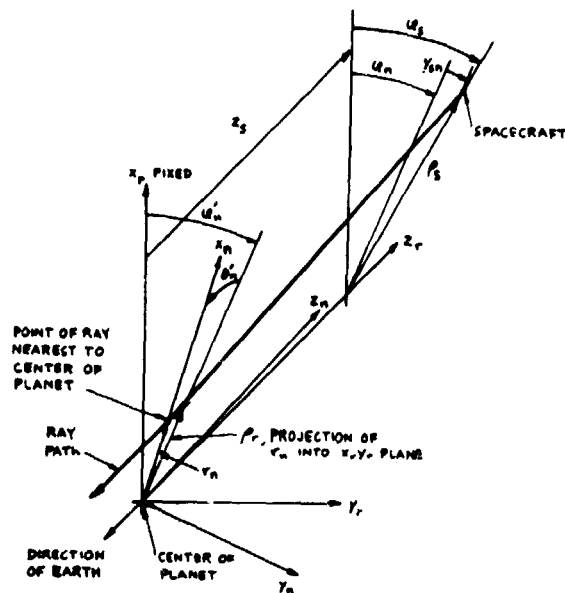


Figure 2.—Occultation geometry. The coordinate system $x_r y_r z_r$ is the system determined by the direction of observation. The $x_n y_n z_n$ coordinate system is determined by the nearest point of a ray and the direction of the ray tangent at the nearest point. The $x_r y_r z_r$ coordinates are fixed during an occultation while the $x_n y_n z_n$ moves with the rays describing the occultation.

The spacecraft position vector $\mathbf{R}(t)$ is designated by the coordinates ρ_s, ϕ_s, z_s . In place of time t , one may represent this position vector parametrically in terms of one of the coordinates. Any convenient one may be used, since in an occultation it is presumed that the position of the spacecraft is known relative to the planet. Here we choose to use ρ_s , though in a circular orbit ϕ_s may be preferred. The choice is not important for the theoretical presentation. Thus, we take $\rho_s, \phi_s = \phi_s(\rho_s)$ and $z_s = z_s(\rho_s)$ as known. The nearest point position vector $r_n(t)$ may be described in terms of its radius r_n and the rotation angles θ'_n, ϕ'_n relative to the ρ_r, ϕ_r, z_r coordinate system. Then differentiation of $\Delta\Phi[r_n(t)] = \Delta\Phi[\mathbf{R}(t)]$ leads to

$$\frac{d}{dt}(\Delta\Phi) = \frac{d}{d\rho_s}(\Delta\Phi) \frac{d\rho_s}{dt} = \frac{\partial}{\partial r_n}(\Delta\Phi) \frac{dr_n}{dt} + \frac{\partial}{\partial \theta'_n}(\Delta\Phi) \frac{d\theta'_n}{dt} + \frac{\partial}{\partial \phi'_n}(\Delta\Phi) \frac{d\phi'_n}{dt} \quad (2)$$

Dividing by dr_n/dt ,

$$\frac{d\rho_s}{dr_n} \frac{d}{d\rho_s}(\Delta\Phi) = \frac{\partial}{\partial r_n}(\Delta\Phi) + \frac{\partial}{\partial \theta'_n}(\Delta\Phi) \frac{d\theta'_n}{dr_n} + \frac{\partial}{\partial \phi'_n}(\Delta\Phi) \frac{d\phi'_n}{dr_n} \quad (3)$$

The function $d/d\rho_s(\Delta\Phi) = d_i/dt(\Delta\Phi)/(d\rho_s/dt)$ is known, since its numerator is obtained from the measured residual Doppler shift and the denominator from the known path of the spacecraft. If $\partial/\partial r_n(\Delta\Phi)$, $\partial/\partial\theta'_n(\Delta\Phi)$ and $\partial/\partial\phi'_n(\Delta\Phi)$ are found as functions of r_n , and if suitable functional relationships are found between the medium's refractivity and the phase path functions, then $\partial N/\partial r_n$, $\partial N/\partial\theta'_n$, $\partial N/\partial\phi'_n$ may be determined from Abel-type transformations as will be shown.

The refractivity is now assumed to be expressed as a Taylor series about some angular coordinates θ_0 , ϕ_0 , where θ_0 is the complement of the usual polar angle in a spherical coordinate system. The dependence on radius is maintained implicitly in all terms. The coordinates θ_0 , ϕ_0 correspond to values in the vicinity of the nearest point of a ray. Then, cut off at the first derivatives,

$$N(r, \theta'_n, \phi'_n) = N(r, \theta_0, \phi_0) + \left. \frac{\partial N(r, \theta, \phi)}{\partial \theta} \right|_{\substack{\theta=\theta_0 \\ \phi=\phi_0}} (\theta'_n - \theta_0) + \left. \frac{\partial N(r, \theta, \phi)}{\partial \phi} \right|_{\substack{\theta=\theta_0 \\ \phi=\phi_0}} (\phi'_n - \phi_0) \quad (4)$$

The derivatives of N from (4) are

$$\frac{\partial N(r, \theta'_n, \phi'_n)}{\partial r} = \frac{\partial N(r, \theta_0, \phi_0)}{\partial r} + \left. \frac{\partial^2 N(r, \theta, \phi)}{\partial r \partial \theta} \right|_{\substack{\theta=\theta_0 \\ \phi=\phi_0}} (\theta'_n - \theta_0) + \left. \frac{\partial^2 N(r, \theta, \phi)}{\partial r \partial \phi} \right|_{\substack{\theta=\theta_0 \\ \phi=\phi_0}} (\phi'_n - \phi_0) \quad (5)$$

$$\frac{\partial N(r, \theta'_n, \phi'_n)}{\partial \theta'_n} = \left. \frac{\partial N(r, \theta, \phi)}{\partial \theta} \right|_{\substack{\theta=\theta_0 \\ \phi=\phi_0}} \quad (6)$$

$$\frac{\partial N(r, \theta'_n, \phi'_n)}{\partial \phi'_n} = \left. \frac{\partial N(r, \theta, \phi)}{\partial \phi} \right|_{\substack{\theta=\theta_0 \\ \phi=\phi_0}} \quad (7)$$

For media in which the angular dependence of the refractivity is small any individual ray with nearest point coordinates r_n , θ'_n , ϕ'_n will be well approximated by a ray in a fictitious medium described by the actual medium along the radial line θ'_n and ϕ'_n equal to constants. Then to first order in ϵ the residual phase length may be expressed by integrating along a straight line tangent to the ray at the nearest point [Pirraglia and Gross, 1970]:

$$\Delta\Phi(r_n, \theta'_n, \phi'_n) = \frac{2\omega\epsilon}{c} \int_0^\infty N(r, \theta'_n, \phi'_n) ds \quad (8)$$

where ω is the frequency, c is the vacuum velocity of light, ds is an element of length along the path and s , the distance along the path measured from the nearest point, is given by $\sqrt{r^2 - r_n^2}$. The factor 2 takes into account the near equal contribution from both sides of the ray path about the nearest point. The upper limit of the integral,

which is the value of s at the sensible limit of the medium, is set equal to infinity, an approximation valid for exponentially decreasing media or refractivity.

Taking derivatives of (8) with respect to θ'_n , ϕ'_n , and r_n , and using (5), (6), and (7),

$$\frac{\partial}{\partial \theta'_n} (\Delta \Phi) = \frac{2\omega\epsilon}{c} \int_0^\infty \left. \frac{\partial N}{\partial \theta} \right|_{\substack{\theta=\theta_0 \\ \phi=\phi_0}} ds \quad (9)$$

$$\frac{\partial}{\partial \phi'_n} (\Delta \Phi) = \frac{2\omega\epsilon}{c} \int_0^\infty \left. \frac{\partial N}{\partial \phi} \right|_{\substack{\theta=\theta_0 \\ \phi=\phi_0}} ds \quad (10)$$

$$\begin{aligned} \frac{\partial}{\partial r_n} (\Delta \Phi) &= \frac{2\omega\epsilon}{c} \int_0^\infty \frac{r_n}{r} \frac{\partial N(r, \theta, \phi)}{\partial r} ds \\ &= \frac{2\omega\epsilon}{c} \int_0^\infty \frac{r_n}{r} \frac{\partial N(r, \theta_0, \phi_0)}{\partial r} ds + (\theta'_n - \theta_0) \frac{d}{dr_n} \left(\frac{\partial \Delta \Phi}{\partial \theta'_n} \right) + (\phi'_n - \phi_0) \frac{d}{dr_n} \left(\frac{\partial \Delta \Phi}{\partial \phi'_n} \right) \end{aligned} \quad (11)$$

If the left-hand sides of (9), (10), and (11) can be found, then the components of the gradient of N may be determined by inversion.

Defining the functions:

$$h(r_n) \equiv \frac{c}{\omega\epsilon} \frac{\partial}{\partial \theta'_n} (\Delta \Phi) \quad (12)$$

$$e(r_n) \equiv \frac{c}{\omega\epsilon} \frac{\partial}{\partial \phi'_n} (\Delta \Phi) \quad (13)$$

$$g(r_n) \equiv 2 \int_0^\infty \frac{r_n}{r} \frac{\partial N(r, \theta_0, \phi_0)}{\partial r} ds \quad (14)$$

$$f(\rho_s) \equiv \frac{c}{\epsilon\omega} \frac{d}{d\rho_s} (\Delta \Phi) \quad (15)$$

(3) may be rewritten

$$\begin{aligned} \frac{d\rho_s}{dr_n} f(\rho_s) = & \epsilon g(r_n) + \epsilon [\theta'_n(r_n) - \theta_0] \frac{dh(r_n)}{dr_n} + \epsilon [\phi'_n(r_n) - \phi_0] \frac{de(r_n)}{dr_n} \\ & + \epsilon \frac{d\theta'_n}{dr_n} h(r_n) + \epsilon \frac{d\phi'_n}{dr_n} e(r_n) \end{aligned} \quad (16)$$

(16) is not sufficient to find g , h , and e , additional equations being needed from the ray path.

Pirraglia and Gross [1970] gave the first-order equation for the ray. With the $\partial N / \partial \theta'_n$ term dropped as of second order in ϵ , it is, in Cartesian coordinates (x_n , y_n , z_n) (fig. 2).

$$r(z_n) = \epsilon \int_0^{z_n} (z_n - \tau) \left[\hat{x}_n \frac{r_n}{r} \frac{\partial N(r, \theta'_n, \phi'_n)}{\partial r} + \hat{y}_n \frac{1}{r_n} \frac{\partial N(r, \theta'_n, \phi'_n)}{\partial \phi'_n} \right] d\tau + r_0 \quad (17)$$

where $r_0 = z_n \hat{z}_n + r_n \hat{x}_n$. The vectors \hat{x}_n , \hat{y}_n , and \hat{z}_n are the unit vectors of the coordinate system.

For large values of $|z_n|$, where the ray is outside the effective medium, the upper limit may be set equal to infinity. With this limit (17) yields straight-line equations representing the asymptotic form of the ray outside the medium:

$$\left. \begin{aligned} x_n &= \epsilon m_x(r_n) z_n - \epsilon c_x(r_n) + r_n \\ y_n &= \epsilon m_y(r_n) z_n - \epsilon c_y(r_n) \end{aligned} \right\} \quad (18)$$

$$\left. \begin{aligned} m_x(r_n) &= \int_0^\infty \frac{r_n}{r} \frac{\partial N}{\partial r} d\tau \\ m_y(r_n) &= \int_0^\infty \frac{1}{r_n} \frac{\partial N}{\partial \phi'_n} d\tau \\ c_x(r_n) &= \int_0^\infty \frac{r_n}{r} \tau \frac{\partial N}{\partial r} d\tau \\ c_y(r_n) &= \int_0^\infty \frac{\tau}{r_n} \frac{\partial N}{\partial \phi'_n} d\tau \end{aligned} \right\} \quad (19)$$

Here, m_x and m_y are the slopes for the ray with nearest radius r_n and position angles θ'_n , ϕ'_n ; c_x and c_y are the constant terms of the straight line equations and may be shown to be related to m by

$$c(r_n) = \frac{2}{\pi} \int_{r_n}^{\infty} \frac{m(x)x dx}{\sqrt{x^2 - r_n^2}} \quad (20)$$

Operationally (20) will be represented by $c = T(m)$.

The ray path, projected onto the $x_n z_n$ and $y_n z_n$ planes is shown in figure 3. From this figure and Eqs. (18) and (19), the spacecraft coordinates may be expressed in terms of the projection of ρ_s on the $x_n z_n$ plane, namely ρ_{sn} , and onto the $y_n z_n$ plane, y_{sn} :

$$\left. \begin{aligned} \rho_{sn} &= 2\epsilon m_x z_n - \epsilon T(m_x) + r_n \\ y_{sn} &= 2\epsilon m_y z_n - \epsilon T(m_y) \end{aligned} \right\} \quad (21)$$

Due to the small bending of the rays the z_n axis and the z_s axis will be rotated from each other by small angles of the order of ϵ , and we may set $\rho_s \approx \rho_{sn}$, and $z_s \approx z_{sn}$, the coordinate of the spacecraft along the z_n axis. From Figure 2 we also see that $y_{sn} \approx \rho_s(\phi_s - \phi_n)$. Also, from the definitions of m_x, m_y, g, h , and e we have

$$\left. \begin{aligned} m_x(r_n) &= \frac{1}{2} \left[g(r_n) + (\theta'_n - \theta_0) \frac{dh(r_n)}{dr_n} \right. \\ &\quad \left. + (\phi'_n - \phi_0) \frac{de(r_n)}{dr_n} \right] \\ m_y(r_n) &= \frac{1}{2r_n} e(r_n) \end{aligned} \right\} \quad (22)$$

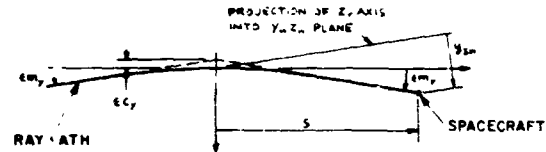
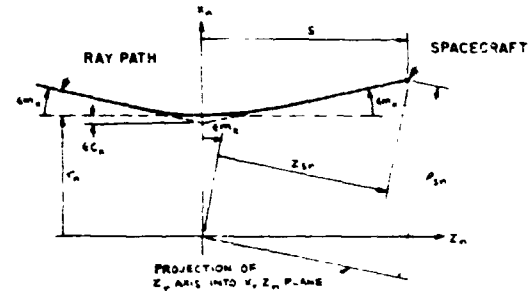


Figure 3. - Projections of ray paths. The upper figure shows the projection of the ray path into the $x_n z_n$ plane. The lower figure shows the projection of the ray into the $y_n z_n$ plane.

so that, from (21) and the approximations associated with the small angle between the z axes,

$$\left. \begin{aligned} \rho_s &= \epsilon z_s G(r_n) - \frac{\epsilon}{2} T \left[G(r_n) \right] + r_n \\ \phi_s &= \phi'_n + \frac{1}{\rho_s} \left\{ \frac{\epsilon z_s}{r_n} e(r_n) - \frac{\epsilon}{2} T \left[\frac{e(r_n)}{r_n} \right] \right\} \end{aligned} \right\} \quad (23)$$

Since bending in the $x_n z_n$ plane equals ϵm_x to first order in ϵ ,

$$\epsilon m_x = \frac{\epsilon}{2} G(r_n) \quad (24)$$

where in (23) and (24)

$$G(r_n) \equiv g(r_n) + (\theta'_n - \theta_o) \frac{dh(r_n)}{dr_n} + (\phi'_n - \phi_o) \frac{de(r_n)}{dr_n} \quad (25)$$

Equations (16), (23), and (25) form the desired set of equations which can be solved if frequency or spatial dispersion is utilized to provide the data. That this is so will be demonstrated for an ionosphere with radiation at two different frequencies. Since ϵ contains the frequency dependent quantity, its value will differ for each frequency. For a given value of r_n for both frequencies the values of ρ_s , z_s , $f(\rho_s)$, ϕ_s , ϕ'_n , and θ'_n will be different. When two frequencies are radiated simultaneously, their corresponding values of r_n will be different, but if the frequencies are close enough or bending of the rays is slight, rays from both frequencies will reach the same r_n at slightly different times. We utilize the z_s and ϕ_s from spacecraft trajectory data as a function of ρ_s , and $f(\rho_s)$ is obtained from the measured Doppler for each frequency and the trajectory data. For a given r_n , the corresponding ρ_s for each frequency is not known. To indicate that quantities are associated with each of two frequencies we utilize the subscripts i , where i takes on the values 1 and 2. Then, since the θ'_n terms contribute to order ϵ^2 , the equations to first order in ϵ from (16) and (23) through (25) become

$$\left. \begin{aligned} \epsilon_i f_i(\rho_{si}) &= \epsilon_i G_i(r_n) + \epsilon_i \frac{d\phi'_{ni}}{d\rho_{si}} e(r_n) \\ \rho_{si} &= \epsilon_i z_{si}(\rho_{si}) G_i(r_n) - \frac{\epsilon_i}{2} T \left[G_i(r_n) \right] + r_n \\ \phi_{si} &= \phi'_{ni}(r_n) + \frac{\epsilon_i}{\rho_{si}} \left\{ z_{si}(\rho_{si}) \frac{e(r_n)}{r_n} - \frac{1}{2} T \left[\frac{e(r_n)}{r_n} \right] \right\} \\ G_i(r_n) &= g(r_n) + \left[\phi'_{ni}(r_n) - \phi_o \right] \frac{de(r_n)}{dr_n} \end{aligned} \right\} \quad (26)$$

For a given r_n , with $i = 1, 2$, we have a set of eight equations in the eight unknowns ρ_{s1} , ρ_{s2} , ϕ'_{n1} , ϕ'_{n2} , G_1 , G_2 , e , and g . Note that to first order in ϵ , $\partial\Phi / \partial\theta'_n$, cannot be determined. By carrying out the expansion to second order in ϵ and utilizing three frequencies, this quantity may, in principle, be determined.

The set of equations (26) may be solved numerically for the functions $g(r_n)$ and $e(r_n)$. From (10), (13), and (14) and $r^2 = r_n^2 + s^2$

$$\left. \begin{aligned} g(r_n) &= 2r_n \int_{r_n}^{\infty} \frac{I}{\sqrt{r^2 - r_n^2}} \frac{\partial N}{\partial r} dr \\ e(r_n) &= 2 \int_{r_n}^{\infty} \frac{(\partial N / \partial \phi'_n) r dr}{\sqrt{r^2 - r_n^2}} \end{aligned} \right\} \quad (27)$$

These equations may be inverted by Abel transforms yielding

$$\left. \begin{aligned} \frac{\partial N}{\partial r}(r, \theta_0, \phi_0) &= -\frac{r}{\pi} \int_0^\infty \left[\frac{g'(\sqrt{q^2 + r^2})}{q^2 + r^2} - \frac{g(\sqrt{q^2 + r^2})}{(q^2 + r^2)^{3/2}} \right] dq \\ \frac{\partial N}{\partial \phi_0}(r, \theta_0, \phi_0) &= -\frac{1}{\pi} \int_0^\infty \frac{e'(\sqrt{q^2 + r^2})}{\sqrt{q^2 + r^2}} dq \end{aligned} \right\} \quad (28)$$

where the prime denotes differentiation with respect to the argument. We see that $\partial N/\partial r$ and $\partial N/\partial \phi_0$ from (28) are the radial and angular components of the gradient transverse to the ray, with θ_0 and ϕ_0 taking on values in the vicinity of the nearest point of each ray.

NUMERICAL CALCULATIONS

Calculations were made for a model of an ionosphere distributed about a planet in which the electron density varies with radius and solar zenith angle and the temperature or scale height varies with this angle. The top of the layer is exponentially decreasing. Below there is a peak in the ionization with a density that varies about the planet but at a fixed altitude. The model is described in terms of a spherical coordinate system (r, θ, ϕ) with the polar axis in the direction from the planet's center to the sun. The origin is at this center, and the polar angle θ is very closely equal to the zenith angle. The medium has axial symmetry in ϕ . At frequencies well above the maximum plasma frequency the index of refraction is given approximately by

$$\mu = 1 - \epsilon \left[d_s - \frac{d_s - d_a}{1 + e^{\nu \cos \theta}} \right] P(R, \theta) \quad (29)$$

Here $\epsilon = 2\pi e^2 n_m / m \omega^2$, n_m is the maximum density at $\theta = \pi/2$ (the terminator), e and m are the electronic charge and mass,

$$P(R, \theta) = \frac{1}{2} \exp - \left[\frac{R - R_m - 2(1 + q \cos \theta)}{1 + q \cos \theta} \right]$$

for the exponential layer, and

$$P(R, \theta) = 1 - \frac{1}{8} \left(\frac{R - R_m}{1 + q \cos \theta} \right)^2$$

for the parabolic peak layer, where $R = (r - r_p)/H_a$ the altitude in scale height units, R_m is the altitude of the level of maximum density in scale height units, H_a is the normalizing scale height and corresponds to the scale height at $\theta = \pi/2$, r is the radius, r_p is the radius of the planet, d_s and d_a are parameters controlling the density level of the entire ionosphere and such that $(d_s + d_a)/2 = 1$, ν is a parameter controlling the rate of change of density with θ , and q is a parameter controlling the distribution of temperature with θ . The subsolar profile is for $\theta = 0$, and

the antisolar profile is for $\theta = \pi$. For large values of ν the ratio of subsolar to antisolar densities at the peaks is equal to d_s/d_a .

For the purpose of numerical calculations we chose the ratio of subsolar to antisolar electron densities as 20:1. The value of q used is 0.5, which gives a scale height ratio of subsolar to antisolar profiles of 3:1. The value of ν was set equal to 10 for three occultations. Ninety percent of the change in electron density occurs within a band of 25° centered at the terminator ($\theta = \pi/2$) for this value of ν . One occultation was calculated for $\nu = 100$ for which 90% of the change occurs in a 3° band. The ratio of the radius at the peak to the scale height was set equal to 105:1. The corresponding radius of the planet is 100. The values of ϵ for the two frequencies were 5×10^{-5} and 1.125×10^{-4} .

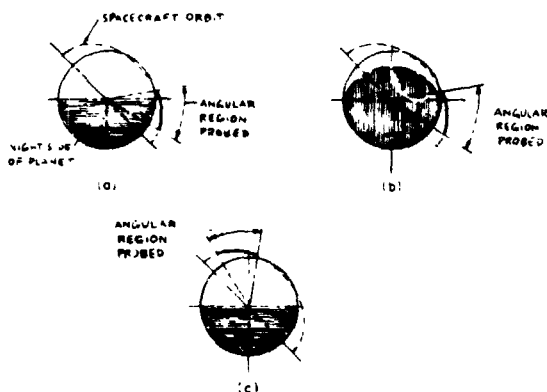


Figure 4. — Occultation orbits. The three figures show the elliptical projection of a circular orbit into the plane perpendicular to the viewing direction. The heavy line indicates the part of the orbit used in the experiment (it is also an approximation to the locus at nearest points). The shaded portions indicate the night-side hemisphere of the planet.

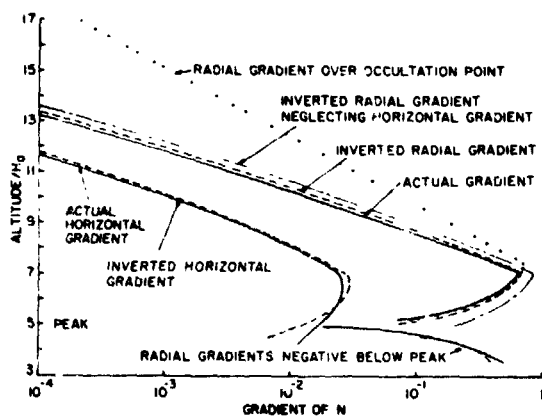


Figure 5. — Gradients of N for occultation indicated in figure 4a. The absolute values of the gradients of N along the locus of nearest points vs. normalized altitude are shown. The radial gradient above the occultation point is also indicated.

If the normalizing scale height is 30 km, the radius to the peak is 3150 km; then the model chosen could be of the scale of the Martian ionosphere. If the plasma frequency at the peak level at $\theta = \pi/2$ is 3.5 MHz, the frequencies associated with the chosen values of ϵ would be 500 and 750 MHz. For the values of ϵ the separation of the rays at their nearest points for the two frequencies never exceeds 0.1 scale height.

The spacecraft is assumed to be in a circular orbit of radius $120 H_a$ and inclined to the direction of the earth, so that its elliptical projection has a semiminor radius of $100 H_a$. The axes of the ellipse are either at 45° to the axis of symmetry or to its projection as viewed from earth for all examples.

The first computation is for $\nu = 10$ and for an occultation starting on the night side. It is viewed perpendicular to the axis of symmetry as shown in figure 4(a). The absolute values of the radial and transverse angular components of the refractivity gradient along the locus of nearest points are shown in figure 5, both for exact values available from the model and computed values deduced by inversion procedures using computed Doppler shift $d\lambda$. The angular component is the horizontal component, transverse to the rays. Radial gradients are shown for inversion with and without corrections for the angular component of the gradient. The actual radial gradient for the profile over the occultation is also plotted. It is seen that if inversion is performed neglecting angular variations and if it is assumed to yield the profile over the occultation point, the interpretation would be considerably in error. The radial gradient errors are much smaller when compared with the actual profile but still are as large as 50% in the exponential region and greater in the parabolic region. Inclusion of the angular variation reduces the error in the radial component by more than a factor of 2. The computed and actual angular components compare very well in the exponential region, but the comparison becomes much poorer below the peak.

If the planet is viewed with the axis of symmetry at an angle of 45° to the plane of projection, as viewed from earth, with the occultation as in figure 4(b), one obtains the graph shown in figure 6. The results are essentially the same as before.

Figure 4(c) indicates an occultation on the day side through a region where the horizontal, or angular, components of the gradient are small. In this case, all the computed radial components of the gradient, the profile along the locus of nearest points, and the profile over the occultation point compare very well (fig. 7). The comparison of horizontal gradients is poor, but the results do indicate that it is smaller than the radial component by a factor of 100.

With $\nu = 100$ and the configuration in figure 4(a) one obtains the results shown in figure 8. The rays pass through a region of large variation of refractivity at $r/H_0 = 108.5$. Two distinct peaks are obtained in the radial component when the angular variation is neglected. Inversion incorporating angular effects considerably improves the results. Although the horizontal gradient from inversion is poor, the large peak is indicated. The results exhibit the difficulties when very large angular gradients are involved. Refinements in numerical procedures may improve the results. The increments used in inversion may be too coarse.

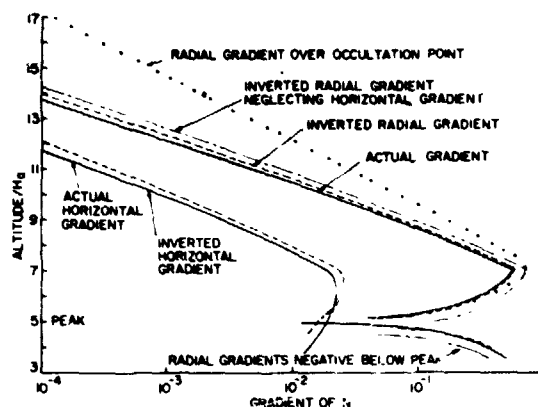


Figure 6.— Gradients of N for occultation indicated in figure 4b. The absolute values of the gradients of N along the locus of nearest points vs. normalized altitude are shown. The radial gradient above the occultation point is also indicated.

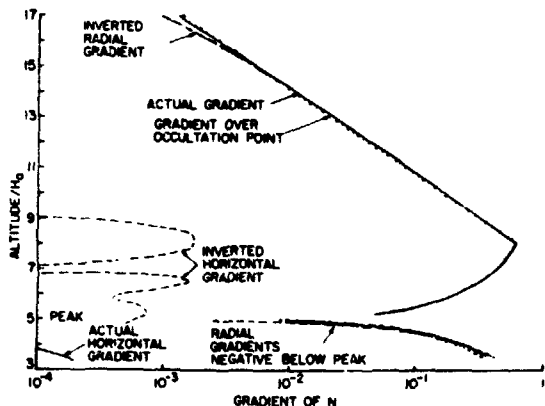


Figure 7.— Gradients of N for occultation indicated in figure 4c. The absolute values of the gradients of N along the locus of nearest points vs. normalized altitude are shown. The radial gradient above the occultation point is also indicated. The actual horizontal gradients are very small and only a small portion is indicated.

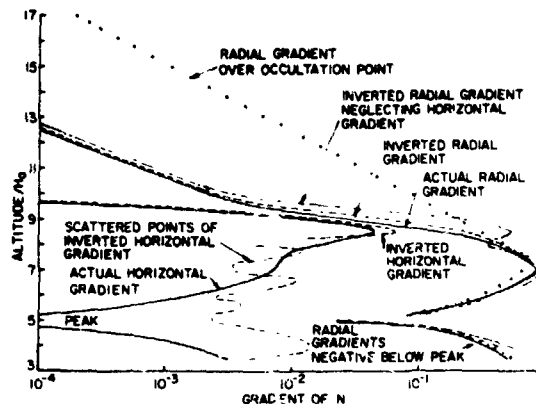


Figure 8.— Gradients of N for occultation indicated in figure 4a and with large angular gradients. The absolute values of gradients of N along the locus of nearest points vs. normalized altitude are shown. The radial gradient above the occultation point is also indicated.

DISCUSSION AND CONCLUSIONS

The feasibility of determining both radial components and angular components of the refractivity transverse to a set of rays in an occultation-type of experiment has been demonstrated. The angular components are found by using frequency dispersion for an ionosphere and spatial dispersion for a neutral atmosphere. The analysis is for slightly refractive media of exponentially limited radial extent in which the angular component of the gradient is not strong, so that a Taylor series in angular coordinates may be cut off at the first derivatives. With the exception of deep atmospheres, it is likely that most regions about a planet will be in accordance with these conditions. Deep atmospheres will have large radial gradients due to their large densities, and the residual Doppler may be mostly from the radial term. The angular contribution may be masked and more difficult to deduce. The physics of deep atmospheres, however, may limit the extent of the angular variations.

The practical feasibility of utilizing dispersion is more limited. Accurate Doppler measurements at more than one frequency would be required for ionization, and dual or multiple payloads would be required for neutral media. The position of the spacecraft about the planet must be accurately known, and the effects of inaccuracies on the determination of angular dependence of the medium requires further study.

The occultation experiment provides information on refractivity of the medium and does not necessarily determine the model of the atmosphere. It could provide scale height information from which temperature could be determined for the assumed mass. However, if the temperature varies with altitude, only the density scale height is determined, which is not simply related to the temperature. The horizontal gradients could provide information useful to studies of possible horizontal motions of the medium. If combined with temperature structure information, even with the recognized uncertainty of composition and scale height interpretation, the measurement of horizontal gradient could serve as a means for understanding the processes controlling the distribution with latitude and longitude. Since the physical theory of atmospheres and ionospheres usually separates vertical and horizontal control of the media, it is difficult to interpret data along a locus of nearest points that is at an angle to either direction without separation of the components of the gradient of the refractivity. The present form of the occultation experiment only approximates the total derivative along the locus which, in general, is different from the radial. Scale heights derived from it could be misleading or useless to interpret, since it would correspond to the variation along the locus. On the other hand, the partial derivatives with respect to radius and angle along the locus would at least separate the radial portion. One can, at least, attempt to approximate a profile along radials to come closer to the density scale heights for the range of angular coverage of the occultation. Measurement of higher angular derivatives would provide better accuracy in restructuring the radial profiles; however, the present practicability appears limited.

Radial components of the gradient of refractivity, as exhibited in figures 5 to 8, may be interpreted in terms of the density scale height along the locus of nearest points, since the data may be graphically analyzed to fit a simple exponential model within a region of altitudes. The graphs of the radial components in figures 5 and 6 appear, at first sight, to be roughly parallel but displaced from each other. Such parallelism implies equal scale heights along the locus with the separation due principally to the magnitude of the refractivity. Closer examination shows that the lines are not quite parallel, implying differences in their scale heights. When the radial gradient as represented by the partial derivatives is combined with the refractivity computed along the locus, the radial scale height may be computed as a function of position along the locus. In general, this will differ substantially from that of the variation of refractivity along the locus and will exhibit the functional dependence on angle. The results for the case of figure 5 are plotted in figure 9.

We have ignored the effects of the interplanetary medium. Gradients may be present in this region that could cause difficulties in establishing the sensible extent of the medium. The practice is to subtract out this contribution. One method utilizes two frequencies [Fjeldbo and Eshleman, 1969] on the basis that the contribution of the interplanetary medium will be the same, except for a frequency scaling factor, for both frequencies. The ray paths to earth will be slightly different for both frequencies, and if gradients are large enough,

the rays for each of the two frequencies could be traveling through regions of different densities. Much more information is needed on the detailed structure of the interplanetary medium to judge this aspect. When two frequencies are utilized to attempt to cancel out the contribution of the interplanetary plasma, then, within the context of this discussion, a third frequency would be necessary to measure the ionospheric radial and angular gradients.

The use of a modified occultation experiment deserves further study. Its use is a compromise between the constraints of spacecraft payload and scientific needs. At best, it cannot compare well with direct probing methods and soundings, methods that require far more spacecraft weight and power.

ACKNOWLEDGMENT

The research reported in this paper was sponsored by the National Aeronautics and Space Administration under Research Grant No. NGR-33-006-047.

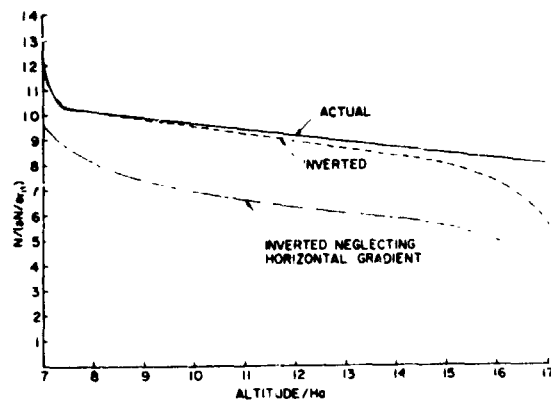


Figure 9.— Relative scale heights derived from the gradients of figure 5. The actual relative scale height is unity.

DISCUSSION

Croft: There is a history here to a closely allied problem in the Russian literature. Gringauz and Alpert had a running commentary from 1965 to 1968. One was measuring the Doppler shift signal from passing satellites. They calculated the integrated electron content as the satellite's altitude changed. He interpreted the data in terms of height profiles. The other author said you couldn't do this because there are localized irregularities having horizontal gradients that would show up as layers in your stratified models. If you start with a single-valued function of one parameter, phase versus time, I don't see how you create two functions of height.

Gross: No, you have two simultaneous measurements. If you can measure for two loci at two frequencies you have two sets of measurements. If the locus of nearest points happen to be tangent to a line of constant refractivity and if you assume spherical symmetry, the derivative with respect to r will be zero at these points. Perhaps, you then interpret this as a peak or a minimum, either in the ionosphere or neutral atmosphere. This could be entirely wrong. When you have the components of the gradient you are then able to separate the information.

Gross: The Russian problem mentioned by Croft is much more difficult because you have the telemeter on the ground and a telemeter on the satellite so, for instance, the basic equation used in occultation where the integral of the Doppler leads you to a phase difference with an integration constant equal to zero is no longer applicable in the Gringauz and Alpert case. Many points of their dispute are therefore not applicable here.

PROFILE INVERSION IN PRESENCE OF RAY BENDING

H. Andrew Wallio

NASA-Langley Research Center

and

Mario D. Grossi

Raytheon Company, Sudbury, Massachusetts

ABSTRACT

N73-11598

Inversion of radio occultation data for planetary atmospheres and ionospheres has been performed by the authors using the seismological Herglotz-Wiechert method, as adapted by Phinney and Anderson to the radio-occultation case. This method does not require the assumption of straightline behavior for the radorays.

Profile reconstruction performed in computer-simulated experiments with this approach have been compared with the ones obtained with the straight-ray Abel transform.

For a thin atmosphere and ionosphere, like the ones encountered on Mars, microwave occultation data can be inverted accurately with both methods. For a dense ionosphere like the sun's corona, ray bending of microwaves is severe, and recovered refractivity by the Herglotz-Wiechert method provides significant improvement over the straight-ray Abel transform: the error reduces from more than 60% to less than 20% at a height of 60×10^3 km above the base of the corona.

INTRODUCTION

Radio occultation measurements of planetary atmospheres and ionospheres have become an integral part of the standard scientific investigations performed by NASA with planetary probes since the mid-1960's, when various spacecraft of the Mariner class performed pioneering radio occultation experiments. These experiments provided the first reliable measures of the profiles of the atmosphere and ionosphere of Mars [Kliore *et al*, 1965; Fjeldbo *et al*, 1966; Fjeldbo and Eshleman, 1968] and of Venus [Kliore *et al*, 1967; Mariner Stanford Group, 1967; Fjeldbo and Eshleman, 1969].

The data are obtained in the form of "Doppler shift residuals" (characterized by errors with $1\sigma = 1$ mm/sec), which are derived by subtracting from the total observed Doppler affecting a 2200-MHz microwave carrier the amount due to the link's geometric changes. The presence of a nonzero residual (for situations in which both link terminals are outside the medium under probing) is indicative of a change in the takeoff or arrival angle of the ray that propagated (oneway or twoway) between earth and spacecraft and traveled, for at least a portion of its path, through the ionosphere and the atmosphere of the planet.

The inversion of the data has been performed with a variety of geometric optics techniques ranging from the closed-form Abel transform (based on the assumptions that the medium possesses spherical symmetry and straight-line behavior for the ray) to model-fitting approaches based on iterative procedures. In principle, for a planet with a thin atmosphere like Mars, all these methods are adequate. For cases like Jupiter or the sun's corona, closed-form inversion algorithms, desirable in many respects above model-fitting approaches, must exclude the straight-line approximation.

One suitable method is available from seismology, based on the Herglotz-Wiechert approach for interpreting seismic data [Herglotz, 1907; Wiechert and Geiger, 1910]. Another method has been published recently by *Fjeldbo et al.* [1971].

In seismology, this bent ray-path technique leads to the derivation of velocity-depth profiles, and to the calculation of the position and the identification of geological features in the earth's mantle. The observables are the travel times of seismic waves between stations located on the earth surface. *Phinney and Anderson* [1968] have shown that this method is usable for the situations in which the observables are the Doppler residuals of the radio occultation measurements. The purpose of our work has been to code their formulas in a computer program and to use it in the reduction and processing of actual radio occultation data.

The seismological approach has been applied to the inversion of simulated occultation data for the sun's corona and Mars. Application to actual experimental data is under way. The inversion accuracies have been compared with the ones that characterize the inversions obtained with the straight-ray Abel transform. We have found that for cases of strong ray bending (as occurs in the microwave occultation probing of the sun's corona), the errors with the Herglotz-Wiechert approach are significantly smaller than with the straight-ray Abel transform. However, for thin atmospheres they are comparable, as expected.

A computer program based on the Phinney and Anderson analysis has been developed in preparation for the reduction of the occultation data to be gathered in 1976 by the Viking spacecraft for Mars and the Sun, and by such missions as the Grand Tour Mission to the outer planets. Meanwhile, the program is being used as a test to reduce cases of past occultation measurements like the one performed by Mariner 4 to Mars.

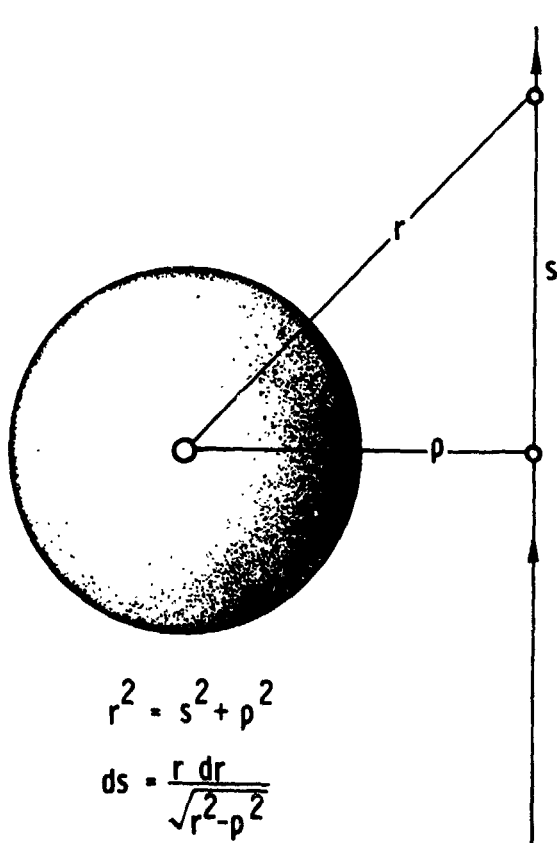


Figure 1. -- Abel transform geometry.

Comparative sensitivity to horizontal gradients and applicability of the seismological approach developed by *Gerver and Markushevich* [1966, 1967] to situations of waveguidance, as encountered for instance in radio occultations occurring at the base of the Venus atmosphere, are under investigation. The applicability of the seismological approach developed by *Backus and Gilbert* [1967, 1968, 1970] is also under study. All these analyses require the computation of the impact parameter of the radio ray from the observed Doppler residuals, a derivation that has been proven feasible by *Phinney and Anderson* [1968].

PROFILE INVERSION BY STRAIGHT-RAY ABEL TRANSFORM

The integration in the time domain of the Doppler residuals provides the differential phase-path length. This quantity is the difference between the straight-line geometrical distance between transmitter and receiver and the phase-path length for radio waves. The integration constant is zero when provisions are made for starting the integration from a position of the probing link completely external to the medium under evaluation.

With reference to figure 1 and assuming a straight-line ray the spherically symmetric medium (index of refraction n is a function of only of r), we obtain the

differential phase-path length

$$\begin{aligned} T(\rho) &= \int_{-\infty}^{+\infty} ds - \int_{-\infty}^{+\infty} n(r) ds \\ &= \int_{-\infty}^{+\infty} [1 - n(r)] ds \end{aligned}$$

This integral is a function of the distance of the ray from the center of the planet at the point of its closest approach to the planet's surface. We call

$$T(\rho) = \int_{-\infty}^{+\infty} [1 - n(r)] ds$$

By changing the variable ds , we can write (fig. 1):

$$T(\rho) = 2 \int_{\rho}^{\infty} \frac{[1 - n(r)] r dr}{\sqrt{r^2 - \rho^2}} = 2 \int_{\rho}^{\infty} \frac{N(r) r dr}{\sqrt{r^2 - \rho^2}} \quad (1)$$

In this equation the radicand in the denominator is nonnegative because for every value of the miss-distance range ρ of the ray from the center of the planet $r > \rho$. Let

$$t = \frac{1}{r^2} \quad r = \frac{1}{\sqrt{t}} \quad r^2 = \frac{1}{t} \quad 2r dr = -\frac{1}{t^2} dt \quad r dr = -\frac{1}{2t^2} dt$$

Equation (1) becomes

$$T(\rho) = \int_{1/\rho^2}^0 \frac{-N[(1/\sqrt{t})] dt}{t^2 \sqrt{(1/t) - \rho^2}} = \int_0^{1/\rho^2} \frac{N[(1/\sqrt{t})] dt}{\rho t^{3/2} \sqrt{\rho^{1/2} - t}} \quad (2)$$

Let

$$x = \frac{1}{\rho^2} \quad \rho = \frac{1}{\sqrt{x}}$$

and we obtain

$$\frac{1}{\sqrt{x}} T\left(\frac{1}{\sqrt{x}}\right) = \int_0^x \frac{N(1/\sqrt{t}) dt}{t^{3/2} \sqrt{x-t}} \quad (3)$$

Applying Abel's integral equation we have,

$$\frac{N(1/\sqrt{t})}{t^{3/2}} = \frac{1}{\pi} \int_0^t \frac{d/dx [(1/\sqrt{x}) T(1/\sqrt{x})]}{\sqrt{t-x}} dx \quad (4)$$

Let

$$t = \frac{1}{r^2} \quad r = \frac{1}{\sqrt{t}} \quad r^2 = \frac{1}{t}$$

We have

$$r^3 N(r) = \frac{1}{\pi} \int_0^{1/r^2} \frac{d/dx [(1/\sqrt{x}) T(1/\sqrt{x})]}{\sqrt{1/r^2 - x}} dx \quad (5)$$

$$\frac{d}{dx} \left[\frac{1}{\sqrt{x}} T\left(\frac{1}{\sqrt{x}}\right) \right] = -\frac{1}{2} \frac{T(1/\sqrt{x})}{x^{3/2}} - \frac{1}{2} \frac{T'(1/\sqrt{x})}{x^2} \quad (6)$$

$$r^3 N(r) = -\frac{1}{2\pi} \int_0^{1/r^2} \frac{[T(1/\sqrt{x})]/x^{3/2} + [T'(1/\sqrt{x})]/x^2}{\sqrt{1/r^2 - x}} dx \quad (7)$$

where $T' = dT(\rho)/d\rho$. Let

$$x = \rho^{1/2} \quad \rho = \frac{1}{\sqrt{x}} \quad 2\rho d\rho = -\frac{1}{x^2} dx$$

We obtain

$$r^3 N(r) = \frac{1}{2\pi} \int_{\infty}^r \frac{\{[T(\rho)/\rho] + T'(\rho)\} 2\rho d\rho}{\sqrt{r^{1/2} - \rho^{1/2}}} \quad (8)$$

$$N(r) = -\frac{1}{\pi^2} \int_r^{\infty} \frac{\rho T(\rho) + \rho^2 T'(\rho)}{\sqrt{\rho^2 - r^2}} d\rho \quad (9)$$

and

$$n(r) = 1 + \frac{1}{\pi^2} \int_r^{\infty} \frac{\rho T(\rho) + \rho^2 T'(\rho)}{\sqrt{\rho^2 - r^2}} d\rho \quad (10)$$

Because of the change in the limits of integration, the radicand in the denominator is nonnegative. In fact, for every radial height r , $N(r)$ and $n(r)$ are obtained from columnar measurements made at miss-distance ρ always larger than r .

The straight-ray Abel transform approach described above has been applied to the Martian ionosphere/atmosphere and the sun's corona, as representative of thin and thick media.

A Hamiltonian ray tracing program was used in conjunction with the chosen models to derive the simulated differential phase-path lengths for various values of ρ .

Figure 2 illustrates the model used for the Mars refractivity profile (MODEL) and the Abel transform reconstruction of this refractivity model (ABEL). Table 1 lists the difference in N units between the model refractivity profile and the Abel transform reconstruction. Figure 3 and table 2 show the same functions for the sun's corona.

It is evident that the Abel transform provides excellent results for the case of the "thin" Martian ionosphere and atmosphere. The errors become intolerable, however, in the reconstruction of the sun's corona when the validity of the straight-line approximation for the ray breaks down.

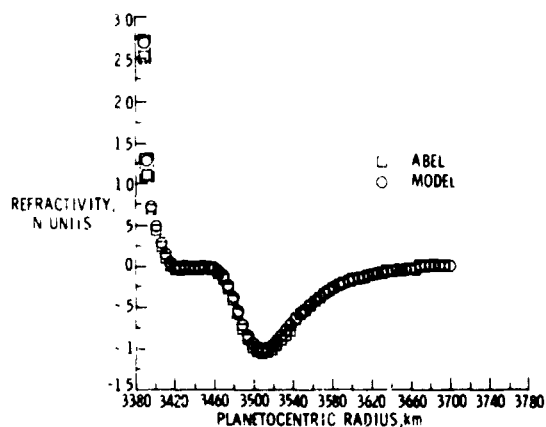


Figure 2.— Model of Mars refractivity profile and its reconstruction by Abel transform (rectangles traced around the pairs of circles and squares identify the original and the reconstructed point).

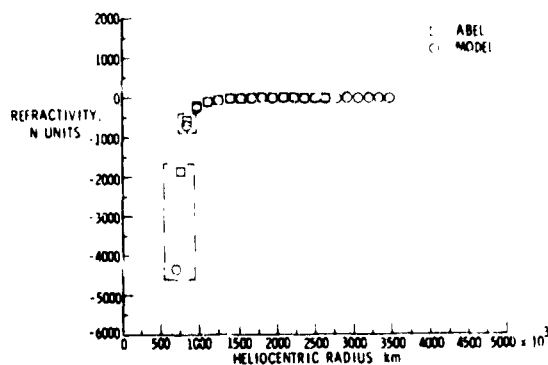


Figure 3.— Model of the Sun corona refractivity profile and its reconstruction by Abel transform (rectangles traced around the pairs of circles and squares identify the original and the reconstructed point).

TABLE 1. ORIGINAL AND RECONSTRUCTED MODEL FOR THE
ATMOSPHERE AND IONOSPHERE OF MARS
(RECONSTRUCTION BY ABEL TRANSFORM)

Propagation Frequency = 2000 MHz

Planetocentric Radius	Differential Phase Delay	Refractivity (Original)	Refractivity (Reconstructed)	Refractivity Difference
3385.8038408	.00049873	2.6910007	2.5343329	-1.5666783E-01
3390.7541166	-.00006561	1.2969379	1.1127441	-1.8419370E-01
3395.6977021	-.00028884	.7342651	.7000476	-3.4217521E-02
3400.6116347	-.00043633	.4823507	.4507898	-3.1560943E-02
3405.5248131	-.00054527	.2818609	.2533993	-2.8462127E-02
3410.4371543	-.00062046	.1327903	.1077764	-2.4814716E-02
3415.3485406	-.00066747	.0351334	.0156015	-2.0131864E-02
3420.2587376	-.00069522	-.0111152	-.0243288	-1.3213628E-02
3425.1669661	-.00070759	-.0059610	-.0112774	-5.3163954E-03
3430.0736711	-.00072737	0.0000000	-.0093382	-9.3381971E-03
3434.9810468	-.00075000	-.0000000	-.0094812	-9.4811747E-03
3439.8883434	-.00077536	-.0000018	-.0095012	-9.4994016E-03
3444.7955183	-.00080421	-.0000454	-.0092206	-9.1752422E-03
3449.7024877	-.00083780	-.0005587	-.0090053	-8.4466883E-03
3454.6091278	-.00087818	-.0038784	-.0115760	-7.6975947E-03
3459.5153911	-.00092821	-.0171971	-.0252309	-8.0737571E-03
3464.4215252	-.00099024	-.0536946	-.0647943	-1.0699684E-02
3469.3281224	-.00106342	-.1274150	-.1446981	-1.7263072E-02
3474.2358337	-.00114185	-.2439253	-.2691008	-2.5175555E-02
3479.1449747	-.00121546	-.3947539	-.4272813	-3.2527304E-02
3484.0553671	-.00129415	-.5601490	-.5978040	-3.7654990E-02
3488.9664814	-.00130720	-.7171498	-.7571699	-4.0020088E-02
3493.8777050	-.00131317	-.8471618	-.8874073	-4.0245454E-02
3498.7885483	-.00129184	-.9396791	-.9789639	-3.9284811E-02
3503.6987291	-.00124699	-.9922043	-1.0303128	-3.8108512E-02
3508.6081505	-.00118405	-1.0080616	-1.0456119	-3.7550312E-02
3513.5167098	-.00110876	-.9937569	-1.0318672	-3.8113433E-02
3518.4257086	-.00102635	-.9568190	-.9969566	-4.0137608E-02
3523.3373833	-.00094117	-.9044097	-.9484911	-4.4081403E-02
3528.250366	-.00085648	-.8426173	-.8930477	-5.0430333E-02
3533.1671808	-.00077461	-.7762285	-.8368103	-6.0581757E-02
3538.0542042	-.00069709	-.7087782	-.7933753	-8.4597098E-02
3542.9611723	-.00062482	-.6427287	-.6389328	3.7959433E-03
3547.8681304	-.00055822	-.5796891	-.5759056	3.7835888E-03
3552.7751089	-.00049739	-.5206237	-.5169908	3.6329555E-03
3557.6821263	-.00044222	-.4660302	-.4626115	3.4187144E-03
3562.5891934	-.00039245	-.4160816	-.4129212	3.1604393E-03
3567.4963147	-.00034772	-.3707338	-.3678395	2.8943039E-03
3572.4034909	-.00030765	-.3298046	-.3271807	2.6238853E-03
3577.3107209	-.00027183	-.2930299	-.2906433	2.5866255E-03
3582.2180017	-.00023988	-.2601030	-.2580032	2.0997693E-03
3587.1253293	-.00021141	-.2307005	-.2287941	1.9063353E-03
3592.0326997	-.00018607	-.2045000	-.2028152	1.6848120E-03
3596.9401086	-.00016352	-.1811908	-.1796952	1.4956071E-03
3601.8475524	-.00014346	-.1604799	-.1591560	1.3739359E-03
3606.7550269	-.00012562	-.1420959	-.1409339	1.177967E-03
3611.6625288	-.00010975	-.1257898	-.1247573	1.024813E-03
3616.5700548	-.00009563	-.1113355	-.1104472	8.87948E-04
3621.4776021	-.00008303	-.0985289	-.0977606	7.6805197E-04
3626.3851679	-.00007179	-.0871856	-.0865069	6.7875790E-04
3631.2927502	-.00006175	-.0771420	-.0765859	5.5616185E-04
3636.2003466	-.00005273	-.0682510	-.0678084	4.4254360E-04
3641.1079552	-.00004459	-.0603816	-.0600339	3.4764660E-04

TABLE 1.- ORIGINAL AND RECONSTRUCTED MODEL FOR THE
ATMOSPHERE AND IONOSPHERE OF MARS
(RECONSTRUCTION BY ABEL TRANSFORM)
(Continued)

Propagation Frequency = 2000 MHz

Planetocentric Radius	Differential Phase Delay	Refractivity (Original)	Refractivity (Reconstructed)	Refractivity Difference
3646.0155746	-.00003720	-.0534174	-.0531748	2.4255063E-04
3650.9232031	-.00003042	-.0472549	-.0472093	4.5665114E-05
3655.8308389	-.00002406	-.0418024	-.0419971	-1.9476145E-04
3660.7384804	-.00001788	-.0369783	-.0378509	-8.7267045E-04
3665.6461239	-.00001127	-.0327104	-.0378314	-5.1210458E-03
3670.5537461	-.00000000	0.0000000	.0000006	5.7527000E-07
3675.4614384	-.00000000	0.0000000	.0000007	6.9534852E-07
3680.3691308	-.00000000	0.0000000	.0000008	8.0721452E-07
3685.2768231	-.00000000	0.0000000	.0000002	1.7881572E-07
3690.1845154	-.00000000	0.0000000	.0000012	1.1589587E-06
3695.0922077	-.00000000	0.0000000	0.0000000	0.

TABLE 2.- ORIGINAL AND RECONSTRUCTED MODEL FOR THE SUN CORONA
(RECONSTRUCTION BY ABEL TRANSFORM)

Propagation Frequency = 2000 MHz

Planetocentric Radius	Differential Phase Delay	Refractivity (Original)	Refractivity (Reconstructed)	Refractivity Difference
7.5786310E+05	-1.47022463E+03	-4.34316152E+03	-1.86435583E+03	2.47880569E+03
8.4348448E+05	-5.84524454E+02	-6.82627600E+02	-5.54988338E+02	1.27539262E+02
9.7715388E+05	-2.45342834E+02	-2.21060135E+02	-1.98717232E+02	2.23429028E+01
1.11509061E+06	-1.22995902E+02	-9.47153676E+01	-9.22099868E+01	2.50538070E+00
1.2538732E+06	-6.78245537E+01	-4.61757491E+01	-6.52852467E+01	-1.91094976E+01
1.3929300E+06	-3.99313143E+01	-2.44567678E+01	-2.27360588E+01	1.67070897E+00
1.5320762E+06	-2.47259222E+01	-1.37900719E+01	-1.29625256E+01	8.27546331E-01
1.6713086E+06	-1.59462278E+01	-8.17848559E+00	-7.74620536E+00	4.32280231E-01
1.8105542E+06	-1.06331003E+01	-5.05856642E+00	-4.82507421E+00	2.33492208E-01
1.9498074E+06	-7.28939278E+00	-3.24259447E+00	-3.11548479E+00	1.27109677E-01
2.0890614E+06	-5.11282164E+00	-2.14543004E+00	-2.07712544E+00	6.63045950E-02
2.2283313E+06	-3.65362588E+00	-1.45520568E+00	-1.42693223E+00	2.82734498E-02
2.3675974E+06	-2.64969763E+00	-1.01145934E+00	-1.01096116E+00	4.98183008E-04
2.5068577E+06	-1.94251251E+00	-7.17816349E-01	-7.44976573E-01	-2.71602249E-02
2.6461533E+06	-1.43321624E+00	-5.18941981E-01	-6.25953850E-01	-1.07011869E-01

THE FEASIBILITY OF DERIVING THE IMPACT PARAMETER FROM THE DOPPLER RESIDUALS

As an immediate consequence of Snell's law, in spherical coordinates, microwave radio rays passing through a refracting medium are characterized by the fact that the product of the index of refraction n , the radius r , and the sine of the zenith angle z is a constant.

In geometric optics treatment of microwave propagation, this constant is of fundamental importance. It is not only the product $n_0 r_0$ at the point of tangency and $r(\sin z)$ at any point beyond the atmosphere, but it is also the radial perpendicular distance to the original undisturbed ray path, or asymptote. It is therefore the necessary parameter to trace a ray through a given atmosphere, and it is a vital quantity in inversion problems.

In particle physics, the perpendicular distance from the center of a force field to the velocity vector of an entering charged particle is called the *impact parameter*. The analogy with the radio rays is crude since the atmospheric refractivity cannot be considered truly equivalent to a force field and our rays are not particles in many respects; nevertheless, "impact parameter" seems to be an appropriate name for the rays' fundamental constant. The exact inversion technique for phase data based on the use of this parameter was presented by *Phinney and Anderson* [1968]. Also, *Tatarsky* [1968a, b] utilized this term as the radius at the tangent point r_0 .

The feasibility of deriving this parameter from the Doppler residuals has been debated in the literature. The analytical relationships between impact parameter and Doppler residuals derived by *Phinney and Anderson* were questioned (incorrectly, in our opinion) by *Graves and Fischbach* [1969], who argued against *Phinney and Anderson's* equations for the eiconal and the directional derivatives in terms of Doppler shifts. These equations are fundamental and give the ray direction with respect to the spacecraft trajectory as a function of the Doppler residuals. *Fischbach* [1970] later reduced the weight of his objectives.

To evaluate the radio ray emission angle from the directional derivative it is necessary that the gradient of the phase or the eiconal equation be determined. We assume an infinite nonconducting isotropic medium where the inductive capacity ϵ is a function of position. Then Maxwell's equations for electric field vector become:

$$\nabla^2 \mathbf{E} + K^2 \mathbf{E} = -\nabla \left(\mathbf{E} \cdot \frac{\nabla \epsilon}{\epsilon} \right) = \nabla^2 \mathbf{E} + K_0^2 n^2 \mathbf{E} \quad (11)$$

where $K_0 = f/c$ is the propagation constant measured in cycles. If we assume that the spatial change of ϵ is small compared to a wavelength

$$\frac{\nabla \epsilon}{\epsilon} \ll \frac{1}{\lambda} \quad (12)$$

then the wave equation reduces to its homogeneous form

$$\nabla^2 \mathbf{E} + K_0^2 n^2 \mathbf{E} = 0 \quad (13)$$

Assume that the form of the solution is:

$$E = A(x, y, z) e^{-iK_0 S(x, y, z)} \quad (14)$$

where $S = S_r + iS_i$ so that

$S_r = \text{constant}$, are surfaces of constant phase

$S_i = \text{constant}$, are surfaces of constant amplitude

Substitution of Eq. (14) into Eq. (13) yields:

$$\nabla^2 \cdot - K_0^2 A \left[n^2 - (\nabla S)^2 \right] - iK_0 \left[A \nabla^2 S + 2(\nabla A)(\nabla S) \right] = 0 \quad (15)$$

If K_0 is large in the sense that

$$\nabla^2 A \ll K_0^2 \quad (16)$$

and

$$\left[A \nabla^2 S + 2(\nabla A)(\nabla S) \right] \ll K_0 \quad (17)$$

then, excluding regions of diffraction, focal points, caustics, and sources, we obtain

$$(\nabla S)^2 = n^2 \quad (18)$$

Now $K_0 S = \phi + ft$ so that

$$(\nabla \phi)^2 = \left(\frac{f}{c} n \right)^2 \quad (19)$$

which is the eiconal equation.

Since n is a function of position, if the spacecraft S in figure 4 is inside the planetary atmosphere, the profile of the atmosphere must be known before Eq. (19) can be solved for $|\nabla \phi|$.

If, however, the spacecraft is above the atmosphere, then n becomes unity and Eq. (19) becomes

$$(\nabla \phi)^2 = \left(\frac{f}{c} \right)^2 \quad (20)$$

The last equation is identical to the eiconal of *Phinney and Anderson* [1968]. Equation (20) is now in a solved form and can be substituted into the equation for the directional derivative to solve for the emission angle.

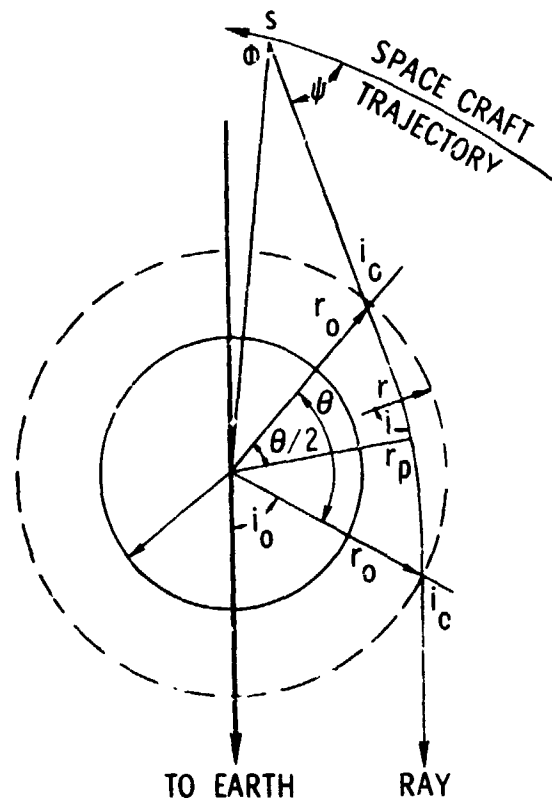


Figure 4.— Geometry of the occultation experiment.

The angle between the ray path and the unperturbed ray path is the deviation angle, and according to our calculation it is

$$\tan \frac{1}{2} (\psi - \psi_e) = \frac{[D_s^2 - (D_e + D_{\text{atm}})^2]^{1/2} - (D_s^2 - D_e^2)^{1/2}}{D_{\text{atm}} + 2D_e} \quad (21)$$

where

D_e = the Doppler shift of the unperturbed ray

D_{atm} = the Doppler residual due to the atmosphere

$D_s = f/c \dot{s}$.

PROFILE INVERSION WITH BENT RAYS BY THE SEISMOLOGICAL APPROACH

The straight-ray approximation can be dropped when referring to inversion approaches like the Herglotz-Wiechert method. In seismology, the observable is the central angle θ subtended by seismic rays in the earth (fig. 4); and the analytical steps are arranged in such a way that the velocity-depth profile is obtained by operating on θ . In radio occultation measurements, the data are obtained in the form of Doppler shift residuals. *Phinney and Anderson* [1968] have shown how the Herglotz-Wiechert approach can be applied to these Doppler residuals, leading to the accurate reconstruction of the refractive index profile, in a condition of ray bending, strong as it may be.

The geometry of the occultation is shown in figure 4. The index of refraction n is a radially dependent function and from Fermat's principle the first variation of the phase-path length should be zero.

$$\delta \int n \, ds = 0 \quad (22)$$

where $ds^2 = dr^2 + r^2 d\theta^2$. The Euler-Lagrange equation that yields a minimum for Eq. (22) is

$$\frac{nr^2 (d\theta/dr)}{[1 + r^2 (d\theta/dr)^2]^{1/2}} = \text{constant} \quad (23)$$

and from the boundary conditions at the minimum radius, Snell's law in spherical geometry is determined. That is

$$p = nr \sin i = n_p r_p = \text{impact parameter} \quad (24)$$

and the impact parameter is a constant for a given ray.

Now let us define the variable η as

$$\eta = nr = \frac{p}{\sin i} \quad (25)$$

From the definition of the path length

$$S = \int ds = \int \left[1 + r^2 \left(\frac{d\theta}{dr} \right)^2 \right]^{1/2} dr \quad (26)$$

and the application of Eqs. (23) through (25), the integral function for the path length is

$$S = 2 \int_{r_p}^r n r (\eta^2 - p^2)^{-1/2} dr \quad (27)$$

The phase (angular measure) along the ray path is

$$\phi = \frac{f}{c} \int n ds \quad (28)$$

and using Eqs. (25) and (27)

$$\phi(p) = 2 \frac{f}{c} \int_{r_p}^{r_0} \frac{\eta^2}{r} (\eta^2 - p^2)^{-1/2} dr \quad (29)$$

we then apply the operator

$$\int_{\eta_1}^{\eta_0} p (p^2 - \eta_1^2)^{-1/2} dp \quad (30)$$

where $\eta_0 \geq \eta_1 \geq \eta_p$, to equation (29). Then, by interchanging the order of integration, integrating by parts repeatedly, and performing a final integration over the appropriate regions, the phase function (29) can be shown to be inverted to

$$r(\eta) = r_0 \exp \left\{ -\frac{c}{\pi f} \int_{\eta_0}^{\eta} \cosh^{-1} \left[\frac{p}{p(\eta)} \right] \frac{1}{p} \frac{d\phi}{dp} dp \right\} \quad (31)$$

Thus the refractivity is found as

$$N(\eta) = \left[\frac{p(\eta)}{r_0 \exp \left\{ -\frac{c}{\pi f} \int_{\eta_0}^{\eta} \cosh^{-1} [p/p(\eta)] \frac{1}{p} \frac{d\phi}{dp} dp \right\}} - 1 \right] \cdot (10^6) \quad (32)$$

Following *Phinney and Anderson* [1968], if we define a ray residual Φ_a

$$\Phi_a = \Phi - \Phi_u \quad (33)$$

where Φ_u is a geometrical term, then

$$N(x) = \left[\exp \left\{ \frac{c}{f\pi} \int_{x_0}^{x_L} \cosh^{-1} \left[\frac{p(x)}{p(x_L)} \right] \frac{1}{p(x)} \frac{d\Phi_a}{dx} dx \right\} - 1 \right] \cdot (10^6) \quad (34)$$

where x is a dummy variable such as time to occultation or satellite position.

In our computer simulated experiment, the impact parameter p is calculated directly from the assumed trajectory and from the Hamiltonian ray tracing program. Therefore, a suitable impact parameter history is known and two forms for the recovered refractivity and minimum probing radius were calculated: Eq. (32), since the data are available from the generating program and it will conform Eq. (31); and Eq. (34), where Φ_a is found in terms of the actual observed phase shift.

Data have been generated for a Martian atmosphere model similar to the one developed from the Mariner 4 results. For this example, ray bending is small (approximately $40 \mu\text{rad}$) and the straight-ray Abel transform has been shown in the past to be a fully adequate approximation. The recovered refractivity from the straight-ray Abel transform (ABEL) and seismic inversion (SEISMIC), and the model refractivity (MODEL) versus geocentric radius of Mars are plotted in figure 5 for the measured phase function.

Table 3 shows that for the thin Martian atmosphere, the errors in recovered refractivity with ABEL and SEISMIC are roughly the same. Also, the minimum probing radius found by the straight-ray Abel transforms is very close to the actual radius, due to the small amount of ray bending.

The Herglotz-Wiechert approach has been applied to the model of the refractivity of the sun's corona. In this example, ray bending becomes significant, and the deviations between the classical straight-ray Abel transform and the seismic approach are clearly noticeable.

A Baumbach electron density model for the sun's corona was assumed and phase data for a frequency of 2,000 MHz were generated. The recovered refractivity profiles from ABEL and SEISMIC are plotted along with the original model (MODEL) versus geocentric radius in figure 6 for the measured phase function.

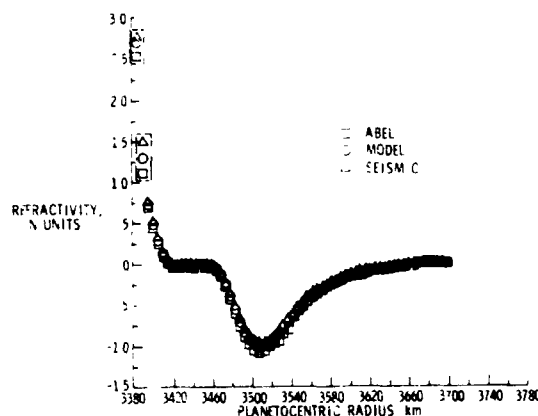


Figure 5.— Model of Mars refractivity profile and its reconstruction by Abel transform and Herglotz-Wiechert inversion (rectangles traced around the pairs of circles, squares, and crosses identify the original and the reconstructed point).

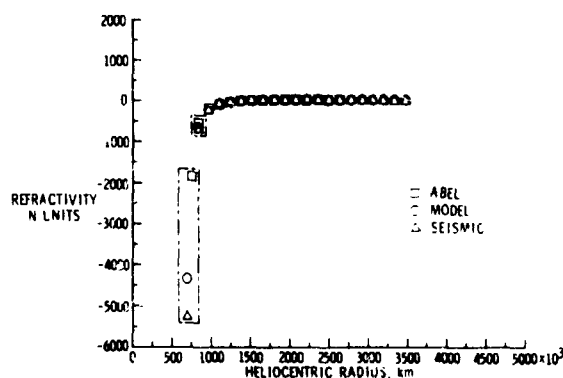


Figure 6.— Model of sun corona refractivity profile and its reconstruction by Abel transform and Herglotz-Wiechert inversion (rectangles traced around the pairs of circles, squares, and crosses identify the original and the reconstructed point).

PLANETOCENTRIC RADIUS	MODEL REFRACTIVITY	ABEL			MERGLITZ-WIECHERT		
		REFRACTIVITY CALCULATED	REFRACTIVITY DIFFERENCE	MINIMUM RADIUS DIFFERENCE	REFRACTIVITY CALCULATED	REFRACTIVITY DIFFERENCE	MINIMUM RADIUS DIFFERENCE
3.385698E+03	2.691001E+00	2.534333E+00	-1.566678E-01	9.452920E-02	2.763462E+00	7.246107E-02	-1.365307E-04
3.390811E+03	1.296930E+00	1.112744E+00	-1.641937E-01	5.666340E-02	1.504153E+00	2.072156E-01	-5.960265E-04
3.395720E+03	7.342651E-01	7.000476E-01	-3.421752E-02	2.267790E-02	7.541950E-01	1.992949E-02	4.000905E-05
3.400629E+03	4.823507E-01	4.507898E-01	-3.156094E-02	1.728530E-02	5.022358E-01	1.988509E-02	4.053678E-05
3.405537E+03	2.818609E-01	2.533988E-01	-2.846213E-02	1.247690E-02	3.236590E-01	4.179815E-02	-3.458183E-05
3.410445E+03	1.327963E-01	1.079754E-01	-2.481482E-02	6.335700E-03	1.523049E-01	1.951459E-02	4.139953E-04
3.415354E+03	3.513338E-02	1.500150E-02	-2.013184E-02	4.979400E-03	5.435087E-02	1.925748E-02	4.104852E-05
3.420261E+03	-1.111522E-02	-2.432880E-02	-1.321363E-02	2.632400E-03	7.502535E-03	1.861775E-02	4.392363E-05
3.425169E+03	-5.961027E-03	-1.127740E-02	-5.316395E-03	2.073400E-03	5.264774E-03	1.122580E-02	6.058667E-05
3.430077E+03	0.	-9.338200E-03	-9.338174E-03	3.048900E-03	1.683858E-02	1.680958E-02	4.942782E-05
3.434984E+03	-2.131628E-08	-9.481200E-03	-9.481174E-03	3.363200E-03	1.676059E-02	1.676062E-02	4.970553E-05
3.439892E+03	-1.769251E-06	-9.501200E-03	-9.499402E-03	3.756600E-03	1.677310E-02	1.677687E-02	4.993872E-05
3.444800E+03	-4.536815E-05	-9.220600E-03	-9.175242E-03	4.271700E-03	1.669785E-02	1.674322E-02	4.490555E-05
3.449707E+03	-5.586571E-04	-9.005300E-03	-8.446688E-03	5.002300E-03	1.62852E-02	1.648718E-02	5.048775E-05
3.454615E+03	-3.875370E-03	-1.157600E-02	-7.697595E-03	6.062270E-03	1.200593E-02	1.588720E-02	5.224672E-05
3.459523E+03	-1.719712E-02	-2.523090E-02	-8.033766E-03	7.538900E-03	-2.375756E-03	1.482137E-02	5.636149E-05
3.464431E+03	-5.369461E-02	-6.479430E-02	-1.109988E-02	9.224800E-03	-3.981830E-03	1.387630E-02	5.937208E-05
3.469339E+03	-1.274150E-01	-1.446981E-01	-1.728307E-02	1.057760E-02	-1.136845E-01	1.373042E-02	5.989768E-05
3.474247E+03	-2.439253E-01	-2.691008E-01	-3.517556E-02	1.095630E-02	-2.291908E-01	1.473450E-02	5.588763E-05
3.479155E+03	-3.947539E-01	-4.272813E-01	-3.252730E-02	1.003530E-02	-3.783475E-01	1.640644E-02	5.027535E-05
3.484063E+03	-5.631490E-01	-5.978040E-01	-3.765499E-02	7.912900E-03	-5.420081E-01	1.814091E-02	4.379454E-05
3.488972E+03	-7.171498E-01	-7.571699E-01	-4.002009E-02	5.048600E-03	-6.977172E-01	1.943255E-02	4.007151E-05
3.493880E+03	-8.471618E-01	-8.874073E-01	-4.024545E-02	1.975000E-03	-8.271119E-01	2.005005E-02	3.711061E-05
3.498788E+03	-9.396791E-01	-9.789639E-01	-3.928481E-02	8.483000E-04	-9.195892E-01	2.008995E-02	3.696965E-05
3.503696E+03	-9.922043E-01	-1.030313E+00	-3.810831E-02	3.149100E-03	-9.724288E-01	1.975145E-02	3.877825E-05
3.508603E+03	-1.008065E+00	-1.045612E+00	-3.755031E-02	4.828500E-03	-9.898412E-01	1.922042E-02	4.056684E-05
3.513511E+03	-9.937569E-01	-1.031867E+00	-3.811034E-02	5.899800E-03	-9.751726E-01	1.858424E-02	4.230456E-05
3.51841E+03	-9.568190E-01	-9.969566E-01	-4.013701E-02	6.468600E-03	-9.347852E-01	1.803383E-02	4.452020E-05
3.523326E+03	-9.044097E-01	-9.484911E-01	-4.408140E-02	6.623300E-03	-8.868474E-01	1.76231E-02	4.534995E-05
3.528234E+03	-8.426173E-01	-8.930476E-01	-5.043033E-02	6.496600E-03	-8.256127E-01	1.720460E-02	4.653235E-05
3.533141E+03	-7.762283E-01	-8.368103E-01	-6.058175E-02	6.180800E-03	-7.592257E-01	1.700278E-02	4.802906E-05
3.538048E+03	-7.087782E-01	-7.933753E-01	-8.459710E-02	5.754200E-03	-6.919622E-01	1.681595E-02	4.799945E-05
3.542956E+03	-6.427287E-01	-6.389328E-01	3.795433E-03	5.252300E-03	-6.260343E-01	1.665442E-02	4.816821E-05
3.547863E+03	-5.796891E-01	-5.759056E-01	3.783589E-03	4.740400E-03	-5.629814E-01	1.670777E-02	4.790372E-05
3.552771E+03	-5.206237E-01	-5.169908E-01	3.632956E-03	4.238900E-03	-5.039402E-01	1.668349E-02	4.822481E-05
3.557678E+03	-4.660302E-01	-4.626115E-01	3.418714E-03	3.756300E-03	-4.449373E-01	1.665659E-02	4.869996E-05
3.562586E+03	-4.160816E-01	-4.129212E-01	3.160439E-03	3.303400E-03	-3.994216E-01	1.66002E-02	4.834261E-05
3.567493E+03	-3.707338E-01	-3.678395E-01	2.894304E-03	2.894700E-03	-3.540226E-01	1.67116E-02	4.779591E-05
3.572401E+03	-3.298046E-01	-3.271807E-01	2.623885E-03	2.520900E-03	-3.130907E-01	1.671389E-02	4.837208E-05

TABLE 3.

PLANETOCENTRIC RADIUS	MODEL REFRACTIVITY	ABEL			HERGLOTZ-WIECHERT		
		REFRACTIVITY CALCULATED	REFRACTIVITY DIFFERENCE	MINIMUM RADIUS DIFFERENCE	REFRACTIVITY CALCULATED	REFRACTIVITY DIFFERENCE	MINIMUM RADIUS DIFFERENCE
3.577309E+03	-2.930299E-01	-2.906433E-01	2.386626E-03	-2.190900E-03	-2.762963E-01	1.673366E-02	4.760509E-05
3.582216E+03	-2.601030E-01	-2.580032E-01	2.099788E-03	-1.891700E-03	-2.432756E-01	1.682739E-02	4.707280E-05
3.587124E+03	-2.307005E-01	-2.287941E-01	1.906335E-03	-1.629300E-03	-2.137700E-01	1.693051E-02	4.687386E-05
3.592031E+03	-2.045000E-01	-2.028152E-01	1.684612E-03	-1.399700E-03	-1.875528E-01	1.694725E-02	4.682249E-05
3.596939E+03	-1.811588E-01	-1.796952E-01	1.495607E-03	-1.198600E-03	-1.642285E-01	1.696726E-02	4.615721E-05
3.601847E+03	-1.604799E-01	-1.591560E-01	1.323936E-03	-1.032400E-03	-1.435270E-01	1.695292E-02	4.675407E-05
3.606754E+03	-1.420959E-01	-1.409339E-01	1.161927E-03	-8.769000E-04	-1.253687E-01	1.702713E-02	4.610226E-05
3.611662E+03	-1.257898E-01	-1.247573E-01	1.032481E-03	-7.388000E-04	-1.096742E-01	1.711555E-02	4.592599E-05
3.616569E+03	-1.111335E-01	-1.104472E-01	8.883185E-04	-6.248000E-04	-9.412870E-02	1.720678E-02	4.502850E-05
3.621477E+03	-9.852864E-02	-9.776060E-02	7.680520E-04	-5.321000E-04	-8.130462E-02	1.722402E-02	4.548815E-05
3.626385E+03	-8.718565E-02	-8.650690E-02	6.787579E-04	-4.673000E-04	-6.992503E-02	1.726062E-02	4.486699E-05
3.631292E+03	-7.714205E-02	-7.658590E-02	5.561519E-04	-3.702000E-04	-6.979477E-02	1.734733E-02	4.477450E-05
3.636200E+03	-6.825098E-02	-6.780840E-02	4.425436E-04	-3.066000E-04	-5.085020E-02	1.740078E-02	4.392184E-05
3.641108E+03	-6.038158E-02	-6.003390E-02	3.476567E-04	-2.552000E-04	-4.287870E-02	1.750286E-02	4.439140E-05
3.646015E+03	-5.341738E-02	-5.317480E-02	2.425506E-04	-2.046000E-04	-3.585352E-02	1.756386E-02	4.331884E-05
3.650923E+03	-4.725492E-02	-4.720930E-02	4.566511E-05	-1.631000E-04	-2.960644E-02	1.764848E-02	4.307326E-05
3.655831E+03	-4.180237E-02	-4.199710E-02	-1.947614E-04	-1.289000E-04	-2.401956E-02	1.778281E-02	4.302399E-05
3.660738E+03	-3.697826E-02	-3.785090E-02	-8.726705E-04	-9.060000E-05	-1.794757E-02	1.794757E-02	4.218853E-05
3.665646E+03	-3.271037E-02	-3.783140E-02	-5.121046E-03	-5.390000E-05	-1.451050E-02	1.819986E-02	4.102734E-05
3.670554E+03	0.	6.000000E-07	5.752700E-07	3.900001E-06	-1.121518E-02	-1.121518E-02	4.129237E-05
3.675461E+03	0.	7.000000E-07	6.953485E-07	1.599998E-06	-1.034117E-02	-1.034117E-02	4.081959E-05
3.680369E+03	0.	8.000000E-07	8.172145E-07	7.999915E-07	-1.066110E-02	-1.066110E-02	3.901715E-05
3.685277E+03	0.	2.000000E-07	1.738157E-07	3.099995E-06	-1.018172E-02	-1.018172E-02	3.759084E-05
3.690185E+03	0.	1.200000E-06	1.158959E-06	4.599991E-06	-9.338130E-03	-9.338130E-03	3.485559E-05
3.695092E+03	0.	0.	0.	2.300003E-06	-6.765603E-03	-6.765603E-03	2.470613E-05

TABLE 3. (continued)

HELIOCENTRIC RADIUS	MODEL REFRACTIVITY	ABEL				MERGLOTZ-WIECHERT			
		REFRACTIVITY CALCULATED	REFRACTIVITY DIFFERENCE	MINIMUM RADIUS DIFFERENCE		REFRACTIVITY CALCULATED	REFRACTIVITY DIFFERENCE	MINIMUM RADIUS DIFFERENCE	
6.99385E+05	-4.343162E+03	-1.864356E+03	2.478806E+03	-5.847763E+04		-5.251392E+03	-9.182301E+02	6.473439E+02	
8.361307E+05	-6.826276E+02	-5.549983E+02	1.276393E+02	-7.293757E+03		-7.473042E+02	-6.467665E+01	5.557484E+01	
9.751027E+05	-2.210661E+02	-1.987172E+02	2.234250E+01	-2.051222E+03		-2.322189E+02	-1.115878E+01	1.212716E+01	
1.114263E+06	-9.471537E+01	-9.220999E+01	2.505381E+00	-8.251700E+02		-9.754703E+01	-2.831661E+00	4.243791E+00	
1.253489E+06	-4.617575E+01	-6.529522E+01	-1.910950E+01	-3.854400E+02		-4.728976E+01	-1.113011E+00	2.361720E+00	
1.392734E+06	-2.445677E+01	-2.278608E+01	1.570759E+00	-1.960800E+02		-2.508665E+01	-6.298858E-01	1.747962E+00	
1.531991E+06	-1.379007E+01	-1.296253E+01	8.275463E-01	-1.059300E+02		-1.443740E+01	-6.475279E-01	1.783167E+00	
1.671247E+06	-8.178436E+00	-7.746205E+00	4.322802E-01	-5.990000E+01		-8.926002E+00	-7.475966E-01	1.974460E+00	
1.810519E+06	-5.058566E+00	-4.825074E+00	2.334922E-01	-3.522000E+01		-5.286475E+00	-2.279088E-01	1.082320E+00	
1.949784E+06	-3.242594E+00	-3.115485E+00	1.271097E-01	-2.127000E+01		-3.640428E+00	-3.978330E-01	1.397464E+00	
2.089048E+06	-2.143430E+00	-2.077125E+00	6.630460E-02	-1.313000E+01		-2.680679E+00	-5.372491E-01	1.702287E+00	
2.228323E+06	-1.455206E+00	-1.426932E+00	2.827345E-02	-8.280000E+00		-1.607415E+00	-1.522089E-01	8.855771E-01	
2.367592E+06	-1.011459E+00	-1.010461E+00	4.981830E-04	-5.240000E+00		-1.313160E+00	-3.017009E-01	1.226672E+00	
2.506854E+06	-7.178163E-01	-7.449764E-01	-2.714022E-02	-3.350000E+00		-1.131507E+00	-4.136904E-01	1.520673E+00	
2.644513E+06	-5.189426E-01	-6.259538E-01	-1.070119E-01	-2.190000E+00		-5.695573E-01	-5.061527E-02	5.917476E-01	

TABLE 4.

As shown in table 4, the divergence of the recovered refractivity and minimum probing radius compared to the original model values increases as the bending becomes appreciable. For the last value calculated, the straight-ray Abel transform was in error 2.47×10^3 N units in the calculated refractivity and 5.84×10^3 km in the minimum probing radius, compared to the seismic error of 0.92×10^3 N units in refractivity and 0.65×10^3 km in minimum probing radius.

EFFECT OF HORIZONTAL GRADIENTS

In the presence of horizontal gradients, the Snell's law equation for the impact parameter, which is used to recover the refractivity, is no longer valid. Therefore, as the severity of the horizontal gradient becomes larger, the Herglotz-Wiechert inversion technique gives poorer reconstruction of the refractivity profile. It is thought that by starting again from Fermat's principle for the new geometry, a better equation for conservation of some parameter for this type of ray may be found. This conservation properly applied to the Herglotz-Wiechert approach may yield better reconstruction of the refractivity profile for horizontal gradients.

CONCLUSIONS AND FUTURE RESEARCH

The work reported in this paper has shown that profile inversion of refractivity data collected with strongly bent rays requires the use of techniques like the seismological Herglotz-Wiechert approach that do not rely on the validity of the straight-line approximation. Inversion errors associated with the application of the straight-ray Abel transform to a medium like the sun's corona have been shown to be intolerably high (larger than 60%).

What still remains are the application of the method to simulated and actual data characterized by instrumentation phase errors and noise; the evaluation of the ability of the algorithm to extract localized averages of the occultation profile from columnar measurements; the evaluation of the profile uniqueness and the related calculation of the probability associated with each one of the "possible" profiles obtained from the data by inversion; the treatment of cases involving waveguidance; and the evaluation of the minimum sampling rate required to assure a desired accuracy.

All of these extensions planned for our research are not expected to invalidate in any respect our present conclusions indicated above. Instead, they will upgrade the capability of the bent-ray approach for the inversion of planetary atmospheres and ionospheres toward the level of sophistication reached through decades of practice by the seismologists in probing the earth's mantle with seismic waves.

ACKNOWLEDGMENTS

The authors thank Dr. William H. Michael, Jr., for his directions and encouragement. In the computational tasks, the authors were assisted by K. A. Smith and G. M. Day.

This research was supported by NASA-Viking Project, Radio Team (Job No. RVK-112 for H. A. W. '10; Contract No. NAS1-9700 to Raytheon Company for M. D. Grossi). The Hamiltonian ray tracing and the straight-ray Abel transform computer programs had been developed under NASA Contract NASW-1772 awarded to Raytheon Co. by NASA Headquarters (Dr. E. R. Schmerling, Project Director).

DISCUSSION

Kliore: Your application of what you call the Abel transform approach labels that transform unfairly as something usable only in the thin atmosphere approximation. Actually you used the Abel transform in both cases. You operate on different functions, one representing a thin atmosphere approximation and the other the general case. What you call the seismological inversion is simply the general case in which you consider everything including the ray bending, etc. It is not clear to me how you obtain the Herglotz-Weichert method. How is that related to the observed phase, which is simply the integral of the Doppler?

Wallio: $\Phi(a)$ is the observed Doppler residual minus a geometric term that is a straight line.

Kliore: That is, Φ is simply the contribution of a phase path through the atmosphere?

Wallio: Φ is actual Doppler along the ray path through the atmosphere, like Phinney and Anderson, but you don't measure that.

Kliore: You seem to imply it was not generally known to use the thin atmosphere approximation; what you call the Abel transform method would not give the correct result in thick atmospheres.

Wallio: I did not mean to imply that.

Gross: The only limitation on the Abel transform, assuming spherical symmetry, is the monotonic variation of the product of the index of refraction μ and the radius r . Mathematically, however, the Abel transform and the seismological technique are essentially the same. Therefore, one would expect to get the same results. Why don't you?

Wallio: In what I called the Abel transform (straight-line approximation), which reduces to what in most books is called Abel's integral equation, the general seismological case reduces to that. If you have bending, however, you are no longer accounting for it in the Abel transform.

Gross: That's not true. Once you assume spherical symmetry the only limitation is the monotonic variation of the product refractive index times the radius. You can have extensive bending, but if you violate this one condition then you have a limitation on the Abel transform.

ERROR ANALYSIS IN THE MARINER 6 AND 7

OCCULTATION EXPERIMENTS

Richard W. Stewart and Joseph S. Hogan

Institute for Space Studies
Goddard Space Flight Center, NASA, New York, New York

N73-11599

ABSTRACT

The Mariner 6 and 7 occultation experiments provide sets of Doppler residuals from which physical properties of the Martian atmosphere may be derived, such as the ionospheric electron density distribution and the lower atmosphere temperature profile. The Doppler residuals contain both systematic and random errors. The former are removed during the data reduction, and the latter are analyzed to yield error limits on the derived physical properties of the atmosphere.

INTRODUCTION

The Mariner series of planetary probes has yielded valuable information on the vertical distribution of temperature and electron density over six points on Mars and two points on Venus. An important aspect of the occultation data analysis is the determination of confidence limits for the physical properties derived from the Doppler residual data. This paper describes a general approach to the question of deriving uncertainties in quantities obtained from the occultation data and applies this approach to the Mariner 6 and 7 experiments. Although the method is general, the derived uncertainties in physical quantities depend on the series of mathematical operations performed on the Doppler residuals to obtain those quantities, and the formulas given in some instances in this paper therefore cannot be used for experiments such as the Mariner 5 Venus occultation where the refraction angle becomes large.

There are two types of errors involved in deriving physical characteristics of the Martian atmosphere from the occultation data: systematic errors due to effects such as oscillator drift and imperfect knowledge of spacecraft trajectory, and random errors due to the noise level of the data.

Systematic errors are manifest in nonzero Doppler residual values at distances sufficiently far from the planet that atmospheric effects cannot be present. We assume that these errors can be removed from the data by subtracting a bias curve, fitted to a subset of pre- or postencounter points, from the entire set of Doppler residuals.

Our analysis of random errors in derived atmospheric characteristics is based on the following assumptions:

1. Systematic errors are removed from the data as stated above. Errors in derived quantities thus reflect only the random errors inherent in the data.
2. Errors in derived quantities are produced only by errors in the Doppler residuals. Uncertainties in other parameters entering into the analysis, such as wavelength of the telemetry signal and spacecraft position, are either negligible compared to errors in the residuals or are systematic and hence removed in the data reduction.
3. The errors in Doppler residuals are statistically independent.

DATA REDUCTION AND ERROR ANALYSIS

Removal of Systematic Error

Ideally, in the pre- and postencounter periods when atmospheric effects are not present in the telemetry signal received from the spacecraft, the Doppler residuals should be scattered about zero—the scatter reflecting the random noise in the data. In practice, systematic errors may be present that produce nonzero residuals in regions far removed from the planet. These systematic errors are removed from the data by performing a polynomial regression analysis on the Doppler residuals over selected time periods in the preencounter (entry) and postencounter (exit) periods. The resultant least-squares fitted bias curve is then extrapolated through the encounter period and subtracted from the complete set of Doppler residuals to obtain a set of corrected residuals.

Figure 1 illustrates this procedure. The dots represent averaged residual data from three stations obtained during Mariner 6 entry; as can be seen, there is a large bias in this particular set of residuals. The curved line through these raw residuals is a third-order polynomial fit to a preencounter set of these points and is extrapolated through the encounter period. The corrected residuals are connected by the solid line running below the raw residuals.

The error involved in extrapolating the bias curve through the encounter period cannot be formally estimated and for this reason an additional constraint must be placed on the data reduction. We assume that there exists a region of essentially zero refractivity in the middle atmosphere of Mars where the electrons are too few to give a negative contribution to the refractivity and the neutral atmosphere too tenuous to give a positive contribution. Most analyses of the corrected residuals do show a region of nearly constant, but not necessarily zero, refractivity in the middle atmosphere; that is, a refractivity ledge occurs in this region. Refractivity ledges that deviate from zero by more than an average of 0.0% refractivity units are assumed to represent cases in which the extrapolation of the bias curve through the encounter period results in large errors in the values of refractivity near the planet. Such cases are rejected, and the procedure of fitting a bias curve is repeated on other sets of residual points until a near-zero refractivity ledge is obtained.

The third-order polynomial fit exhibited in figure 1 is not a best fit in the sense of minimizing the sum of the squared deviations of the fitted points from the bias curve. Best fits in this sense are obtained with higher order polynomials, but these polynomials are erratic when extrapolated outside the region of fit. A change of a single point in the set selected for the polynomial fit can result in radically different extrapolated values outside the region of fit for higher order polynomials, but the third order polynomials are fairly stable to such changes.

We must emphasize that this procedure for correcting systematic errors in the data is an assumption and that the effects of such errors on the results cannot be rigorously analyzed. We believe their effect can be estimated however, by reducing data for a large number of near-zero ledge cases obtained after subtracting a bias curve fitted to several different sets of residual points in the pre- and postencounter periods. Accordingly, the results for the various atmospheric profiles presented in this paper are averages resulting from analyses of 300 near-zero ledge cases

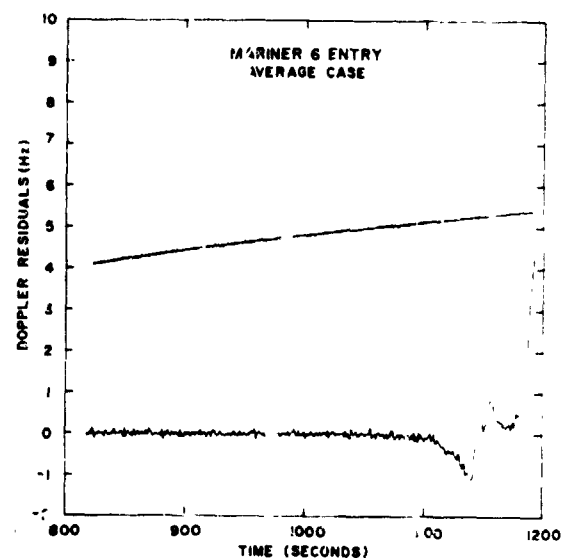


Figure 1.— Raw residuals, polynomial fit, and corrected residuals for Mariner 6 entry.

for each entry and exit. The standard deviations in these average results, when large enough, are indicated in figures by a shaded region about the average profile, while typical standard deviations due to random errors are indicated by error bars.

General Procedure for the Determination of Random Error

The polynomial regression analysis described above yields the variance of the Doppler residual points to which the bias curve is fitted. Since the physical characteristics of the Martian atmosphere are derived from these residuals, the initial error will propagate through the calculations and will ultimately appear as error bars on the various atmospheric profiles. We denote the variance in the Doppler residuals at the j th level in the atmosphere by σ_j^D , where the superscript D refers to Doppler residuals. The variance in a derived quantity Q at the n th level in the atmosphere is then [Beers, 1957].

$$\sigma_n^Q = \sum_{j=1}^n \left(\frac{\partial Q_i}{\partial D_j} \right)^2 \sigma_j^D \quad (1)$$

where Q_n is the value of Q at the n th level and is assumed to be a function of the Doppler residuals at altitude n and at all higher altitudes—that is, at all levels $j \leq n$.

It is convenient to define two matrices M and S by

$$M_{ij}^{QU} = \frac{\partial Q_i}{\partial U_j} \quad S_{ij}^{QU} = \left(\frac{\partial Q_i}{\partial U_j} \right)^2 = \left(M_{ij}^{QU} \right)^2 \quad (2)$$

That is, the j th element of the matrix M^{QU} is the derivative of Q at altitude i with respect to U at altitude j ; Q and U are any two general quantities derived from (or equal to) the Doppler residuals. The matrix S^{QU} is obtained by squaring the individual elements of M^{QU} . Since all derived quantities are obtained by downward integration from the top of the atmosphere, their values at a given altitude do not depend on the residuals at lower altitudes. With a single exception noted later, all the matrices used in this analysis are therefore either lower triangular or diagonal. The specific methods used in obtaining the random error in the various atmospheric quantities derived from the Doppler residuals are enumerated below.

Phase Path

The phase path is computed by time integration of the Doppler residuals according to the trapezoidal rule:

$$P_i = \left(\frac{D_1 + D_2}{2} \right) \Delta t_{1,2} + \left(\frac{D_2 + D_3}{2} \right) \Delta t_{2,3} + \dots + \left(\frac{D_{i-1} + D_i}{2} \right) \Delta t_{i-1,i} \quad i > 1 \quad (3)$$

where $\Delta t_{i-1,i}$ is the time interval between the residual measurements at times t_{i-1} and t_i . The matrix elements involved in calculation of the errors in the phase path increase as

$$\begin{aligned}
M_{ij}^{PD} &= 0 & i = 1 \text{ or } i < 1 \\
&= \frac{1}{2} \Delta t_{j,j+1} & i > 1 \text{ and } j = 1 \\
&= \frac{1}{2} (\Delta t_{j-1,j} + \Delta t_{j,j+1}) & 1 < j < i \\
&= \frac{1}{2} \Delta t_{j-1,j} & j = i > 1
\end{aligned} \tag{4}$$

The phase path variance is

$$\sigma^P = S^{PD} \cdot \sigma^D \tag{5}$$

where σ^P is the vector whose components are the phase path variances at points j , σ^D is the (constant) vector of the residuals variance, and S^{PD} is defined by Eqs. (2) and (4). If we assume a constant Δt in Eq. (3) then it is easy to show that the standard deviation in the phase points increases with decreasing altitude approximately as \sqrt{n} where n is the number of quadrature points.

The magnitude of the errors in phase path resulting from the random error in the Mariner 6 entry residuals in figure 1 is shown by the error bars in figure 2. The maximum standard deviation in phase path is about 0.63 cycles (out of ~8 cycles) near the ground. The shaded region in figure 2 shows the magnitude of the standard deviation in average phase path (given by the solid line) resulting from the reduction of 300 cases. The maximum value of this standard deviation is about 0.49 cycles and occurs near the ground.

Straight-line Approximation to Phase

The straight-line approximation to the phase path P' is [Fjeldbo and Eshleman, 1968].

$$P'_i = P_i - \frac{Z_i}{2\lambda} \alpha_i^2 \tag{6}$$

where Z_i is the distance of the spacecraft behind the planet, λ is the wavelength of the telemetry signal, and α_i is the angle through which the ray is bent in traversing the atmosphere. As before, the subscript i denotes the i th altitude. The variance in P' is given by

$$\sigma^{P'} = S^{P'D} \cdot \sigma^D \tag{7}$$

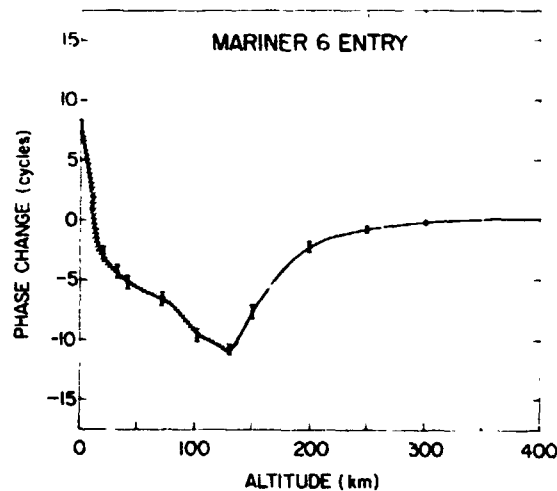


Figure 2.- Average phase path for Mariner 6 entry obtained from three hundred reductions of Doppler residual data, using different bias curves. The standard deviation in the average is indicated by the shaded area. The error bars indicate the magnitude of the standard deviation due to random error derived in any single reduction of the data.

where $\sigma^{P'}$ is the vector whose components are the variances of P' at levels i . The matrix $S^{P'D}$ is computed using (2) once $M^{P'D}$ is computed from the equation

$$M^{P'D} = M^{P'P} \cdot M^{PD} + M^{P'\alpha} \cdot M^{\alpha D} \quad (8)$$

The matrices $M^{P'P}$ and $M^{P'\alpha}$ are calculated from the definition (2) using Eq. (6) to relate P' and P and α . M^{PD} has already been computed in deriving the phase path errors, and $M^{\alpha D}$ can be computed from the relation

$$\alpha_i = - \frac{\lambda D_i}{V_i} \quad (9)$$

where V_i is the component of the spacecraft velocity along the perpendicular to the Earth-Mars line at altitude i . The matrix $M^{P'P}$ in the first term on the right-hand side of (8) is the unit matrix while both $M^{P'\alpha}$ and $M^{\alpha D}$ are diagonal and hence their product is diagonal. The matrix $M^{P'D}$ then differs from M^{PD} only in the diagonal elements which are given by

$$M_{ii}^{P'D} = \frac{1}{2} \Delta t_{i-1,i} + \frac{Z_i \alpha_i}{V_i} \quad i > 1 \quad (10)$$

The errors in P' are only slightly greater than those in P and will not be shown explicitly.

Refractivity

The variance in refractivity is given by

$$\sigma^R = S^{RD} \cdot \sigma^D \quad (11)$$

The refractivities are not given explicitly as functions of the Doppler residuals, but are given in terms of the straight-line approximation to the phase path by [Fjeldbo and Eshleman, 1968].

$$R_1 = \frac{10^6 \lambda P'_1}{\Delta Z_{11}}$$

$$R_i = \frac{10^6 \lambda P'_i - 2 \sum_{j=1}^{i-1} R_j \Delta Z_{ij}}{\Delta Z_{ii}} \quad (12)$$

where ΔZ_{ij} is one-half the distance traversed by the i th ray in the j th layer. The matrix M^{RD} required for the computation of S^{RD} may therefore be obtained from

$$M^{RD} = M^{RP'} \cdot M^{P'D} \quad (13)$$

where $M^{P'D}$ has already been computed in the analysis of P' errors and $M^{RP'}$ can be calculated from the recursion formulae (12). If we define a matrix Z by

$$Z_{ii} = \frac{10^6 \lambda}{\Delta Z_{ii}} \quad Z_{ij} = -2 \frac{\Delta Z_{ij}}{\Delta Z_{ii}} \quad i > j \quad (14)$$

the the matrix $M^{RP'}$ is computed from

$$M_{ik}^{RP'} = Z_{ii} \quad (15)$$

and

$$M_{ik}^{RP'} = \sum_{j=1}^{i-1} Z_{ij} M_{jk}^{RP'} \quad i > k \quad (16)$$

The standard deviations in refractivity are shown by the error bars in figure 3 for the Mariner 6 entry case. These standard deviations, like those in phase, increase slowly with decreasing altitude. The maximum refractivity error occurs near the ground and has a magnitude of about 0.068 refractivity units. Near the electron density peak at 136 km, the standard deviation in refractivity is about 0.046. The standard deviation in average refractivity resulting from the reduction of 300 cases cannot be readily shown on the scale of figure 3. Near the ground, this standard deviation in average refractivity is 0.039 units and decreases to 0.016 units near the electron density maximum.

Electron Density

The basic equation for the variance in electron density is

$$\sigma^N = S^{ND} \cdot \sigma^D \quad (17)$$

The matrix M^{ND} required for the computation of S^{ND} is obtained from

$$M^{ND} = M^{NR} \cdot M^{RD} \quad (18)$$

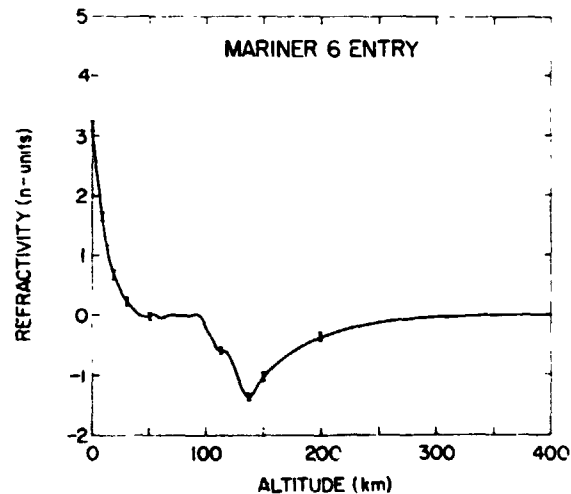


Figure 3.— Average refractivity for Mariner 6 entry obtained from three hundred reductions of Doppler residual data using different bias curves. The standard deviation in the average is too small to be shown on this scale. The error bars indicate the magnitude of the standard deviation due to random error derived in any single reduction of the data.

We computed M^{RD} in the analysis of the refractivity errors, and M^{NR} is a multiple of the unit matrix, which is readily calculated from the relationship between electron density and refractivity

$$N_i = \frac{K_{ij} f^2}{40.3 \times 10^{12}} \text{ (cm}^{-3}\text{)} \quad (19)$$

where f is the frequency of the telemetry signal in Hz. The matrix M^{NR} is thus

$$M_{ij}^{NR} = - \frac{f^2}{40.3 \times 10^{12}} \delta_{ij} \quad (20)$$

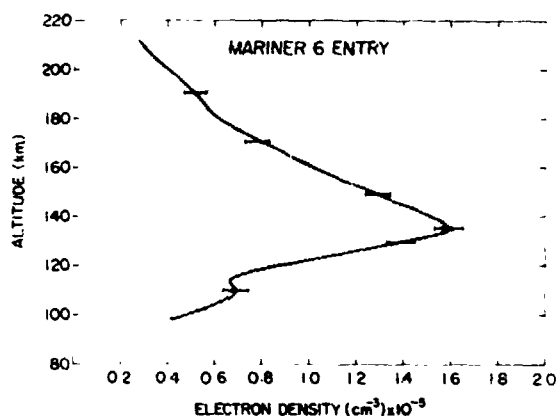


Figure 4.— Average electron density for Mariner 6 entry obtained from three hundred reductions of Doppler residual data using different bias curves. The standard deviation in the average is indicated by the shaded area. The error bars indicate the magnitude of the standard deviation due to random error derived in any single reduction of the data

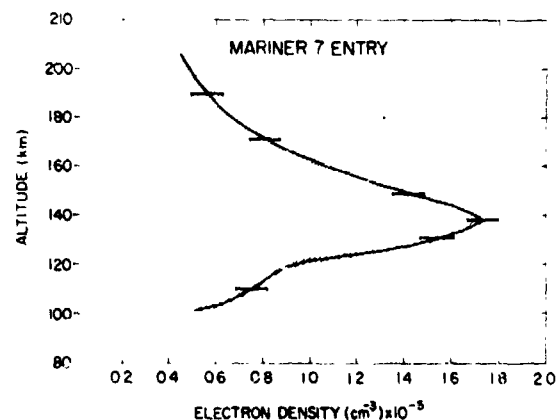


Figure 5.— Same as figure 4 for Mariner 7 entry.

where δ_{ij} is the Kronecker delta. The standard deviation in electron density for Mariners 6 and 7 are shown by the error bars in figures 4 and 5. Since the electron density is proportional to refractivity these standard deviations increase slowly with decreasing altitude. The shaded areas in figures 4 and 5 give the standard deviation in average electron density resulting from 300 reductions of the data.

Temperature

The temperature variance is given by

$$\sigma^T = S^{TD} \cdot \sigma^D \quad (21)$$

The matrix M^{TD} required for the computation of S^{TD} is obtained from

$$M^{TD} = M^{TR} \cdot M^{RD} \quad (22)$$

where M^{RD} has been calculated in the analysis of the refractivity errors, and M^{TR} is calculated from the equations relating temperature to refractivity. Since the temperature at the top of the lower atmosphere, say T_1 , is computed from the scale height determined by the two uppermost refractivity points (the two points just below the ledge), the matrix M^{TR} has a single nonzero element above the diagonal and is the exception noted previously to the set of diagonal and lower triangular matrices. The matrix elements at the initial altitude are

$$M_{11}^{TR} = \frac{\bar{m}\bar{g}}{k} \Delta h_1 \cdot \frac{1}{R_1} \ln\left(\frac{R_2}{R_1}\right)^{-1} \quad (23)$$

$$M_{12}^{TR} = -\frac{R_1}{R_2} M_{11}^{TR} \quad (24)$$

where \bar{m} is the assumed mean molecular mass, \bar{g} is the average acceleration of gravity in the uppermost layer Δh_{12} , and k is Boltzmann's constant. The temperature distribution in the lower atmosphere is related to refractivity by [Fjeldbo and Eshleman, 1968].

$$T(h) = T(h_0) \frac{R(h_0)}{R(h)} + \frac{1}{kR(h)} \int_h^{h_0} g(\xi) \bar{m}(\xi) R(\xi) d\xi \quad (25)$$

h_0 refers to the initial altitude and h to some altitude below h_0 . If we use the trapezoidal rule to integrate Eq. (25) then the diagonal matrix elements of M^{TR} are

$$M_{ii}^{TR} = -\frac{T(h_0) R(h_0)}{R_i^2} + \frac{\bar{m}}{kR_i} \left[\frac{g_i}{2} \Delta h_{i-1,i} - \frac{1}{R_i} X \right] \quad (26)$$

where

$$X = (g_1 R_1 + g_2 R_2) \frac{\Delta h_{12}}{2} + \dots + (g_{i-1} R_{i-1} + g_i R_i) \frac{\Delta h_{i-1,i}}{2} \quad (27)$$

and the off-diagonal elements are

$$M_{ij}^{TR} = \frac{\bar{m}}{2kR_i} \left[g_{j-1} \Delta h_{j-1,j} + g_j \Delta h_{j,j+1} \right] \quad i > j \quad (28)$$

The standard deviations in temperature are shown by the error bars in figures 6 and 7. The shaded areas in figures 6 and 7 give the standard deviation in average temperature resulting from 300 data reductions.

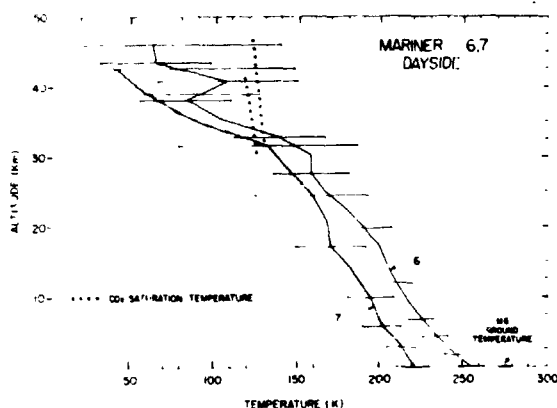


Figure 6.— Average dayside temperature profiles for Mariner 6 and 7 obtained from three hundred reductions of Doppler residual data using different bias curves. The standard deviations in these averages are indicated by the shaded areas where it is large enough to be clearly shown. The error bars indicate the magnitudes of the standard deviations due to random errors derived in any single reduction of the data.

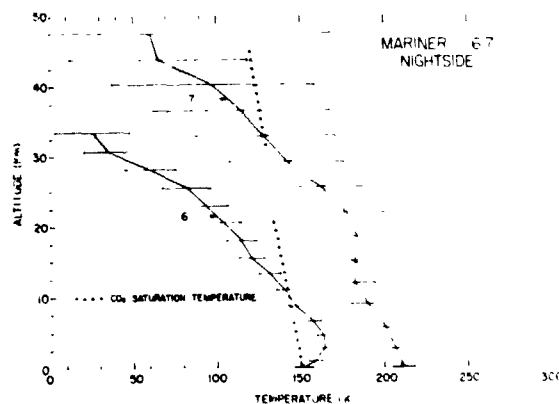


Figure 7.— Same as figure 6 for the Mariner 6 and 7 nightside temperatures.

DISCUSSION

The analysis of the propagation of random error in the Doppler residuals through the data-reduction procedure shows that physical properties of the Martian atmosphere such as the vertical distributions of electron density and lower atmosphere temperature can be determined with good accuracy. The standard deviation shown by the error bars in figures 2 through 7 are properly regarded as lower limits to the actual error present in these quantities since the effect of systematic errors cannot be rigorously analyzed. We have estimated the effect of systematic errors by analyzing a large number of near-zero ledge cases obtained after subtracting a bias curve fitted to several different sets of residual points in the pre- and postencounter periods.

Generally, that standard deviation in the average profile for these cases is less than the standard deviation in any one profile due to random errors in the residuals. The bias removal technique employed in the data reduction thus gives reliable results as long as the constraint of a small refractivity ledge in the middle atmosphere is imposed and it appears that the effect of systematic errors can be nearly eliminated by subtracting a least squares fitted bias curve from the residual data. We cannot prove however, within the confines of the occultation experiment itself, that our results are valid when systematic errors are present, but if they are assumed valid then the random error limits should provide a realistic estimate of the actual uncertainty in physical properties derived from the Doppler residuals.

ACKNOWLEDGMENTS

The authors wish to extend their thanks to Dr. A. Kliore for providing both the Doppler residual data and several useful discussions regarding data-reduction procedures. Dr. L. Russell was of assistance in programming the calculations.

DISCUSSION

Kliore: This is an excellent piece of work. However, another source of systematic error is the fact that you have to choose where to remove the refractivity bias once you have obtained the refractivity curve. I found that that does make a significant difference in the surface conditions of pressure and temperature. But the temperatures are still within your error bars.

Stewart: In this procedure, if you insist on getting a zero ledge in the first place then we really have to remove a refractivity bias before you analyze the data. Of course, you never get to zero, but you can get down to 10^{-4} .

ERROR ANALYSIS OF BENT-RAY RADIO OCCULTATION MEASUREMENTS

R. L. St. Germain

Raytheon Company, Sudbury, Massachusetts

N73-11600

The radio occultation method of studying planetary atmospheres has been used successfully to study the atmospheres of several planets. Radial refractivity profiles of the Martian atmosphere have been achieved by Mariners 4, 6 and 7 and of the Venusian atmosphere by Mariner 5 [Fjeldbo and Eshleman, 1968, Fjeldbo et al., 1970; Kliore et al., 1965; Kliore et al., 1970]. The Mariner '71 orbiter is expected to provide up to several score of occultation data at various points of the Martian surface, and the Viking '75 dual orbiter to Mars will upgrade the spatial coverage and measurement accuracy even more.

There are two types of experimental measurement errors of the Doppler data associated with the radio occultation, random and systematic. Random errors are due to thermal noise in the transmission channel, and the phase lock loop, and quantization error in the digital circuitry. These are called noise type errors. The systematic errors are due to geometric uncertainty and equipment phase instability. It is assumed that a sufficient correction may be made to render the systematic errors small in comparison to the random errors.

This paper concerns the amount of uncertainty, due to random measurement errors, in the refractivity profiles reconstructed by this type of indirect sensing experiment. A class of refractivity profiles is defined which approximately (within a certain uncertainty) fit the set of measured data. Bounds are placed on the extent of this class of solution profiles. To accomplish this, we must examine (1) the sensitivity of the reconstructed refractivity profiles to errors in the measured quantity, that is, the stability of the inversion process to measurement errors; and (2) the statistics of the errors in the measurement.

We present first a brief outline of the direct problem. Then we introduce a linearized Taylor expansion about a given atmospheric model to determine the response of the inversion operator to perturbations of the Doppler data. Numerical calculations of the rms refractivity error (standard deviation) are presented for typical planetary atmospheres. A region of confidence of the reconstructed profiles is related to the standard deviation of the Doppler error.

In the radio occultation experiment, the ray between the spacecraft and the earth tracking station is refracted and retarded as the ray path intersects the planetary atmosphere. This change in phase-path length introduces a Doppler shift. With the positions of the spacecraft, planet under study, and earth known, the geometric Doppler due to their relative motion may be subtracted from the measured Doppler. The medium-induced Doppler frequency residual is the input to an inversion scheme that computes, as the end product, a radial refractivity profile. The direct problem involves integrating the differential phase path length over the ray path. The resulting phase shift (in cycles) is

$$\Phi(p) = 2k_0 \int_{r_p}^{r_0} \frac{\eta}{r} (\eta^2 - p^2)^{1/2} dr \quad (1)$$

where

- $k_0 = f/c$
- $r_0 =$ radius of maximum extent of the planetary atmosphere
- $r_p =$ closest approach radius of the ray path

The ray parameter is a characteristic of each ray and is defined as

$$p = nr \sin i$$

Let

$$\eta = nr$$

$$n = \text{index of refraction}$$

$$r = \text{planetocentric radius}$$

$$i = \text{angle defined in figure 1.}$$

In all practical situations, only a finite amount of data exists, and therefore the Doppler record is uniquely determined by its projection onto a finite dimensional space. Consider a metric space D containing all possible Doppler residuals and another metric space R containing all possible radial refractivity profiles. Every Doppler record and refractivity profile is represented by a vector in the corresponding space. Equation (1) defines a transformation operator that relates each atmospheric model vector to a Doppler record vector (fig. 2); that is,

$$\Phi = \sigma(n) \quad (2)$$

All real data are contaminated by measurement errors. We define the error vectors in the Doppler phase residual and model parameter spaces by

$$\epsilon_\phi \equiv \phi - \phi_0 \quad \text{and} \quad \epsilon_n \equiv n - n_0 \quad (3)$$

where Φ and n are the measured vectors in the Doppler phase and refractivity model spaces, respectively. The subscript 0 represents the true (actual) quantities. Note that

$$\Phi_0 = \sigma(n_0) \quad \text{and} \quad \Phi = \sigma(n) \quad (4)$$

We may expand each component of the refractivity profile vector n by Taylor's formula for a scalar-valued function of several variables. Grouping the terms in matrix form and applying Eq. (3) yields a multidimensional Taylor series (expansion of a

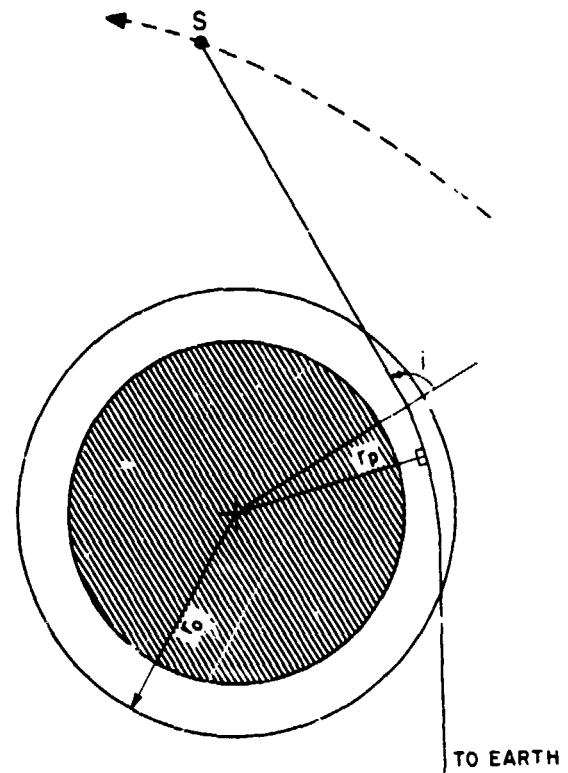


Figure 1.— Ray geometry.

INDIRECTLY SENSED
QUANTITY

MEASURED QUANTITY

STRUCTURE (MODEL)
PARAMETER SPACE

DOPPLER PARAMETER
SPACE

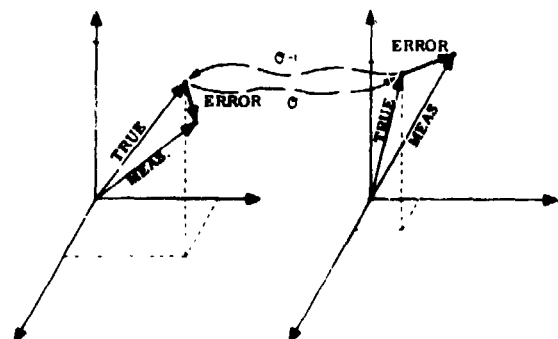


Figure 2.— Transformation between the atmosphere's model vector and Doppler record vector.

vector-valued function of several variables) about a given atmospheric refractivity model:

$$\epsilon_n = \bar{K}_1 \epsilon_\phi + \dots \text{higher order terms} \quad (5)$$

where

$$K_{ij} = \left. \frac{\partial n_i}{\partial \phi_j} \right|_{\phi_0} \quad (6)$$

The response to perturbation (errors) can therefore be linearized about ϕ_0 by neglecting quadratic and high order terms because ϵ_ϕ is small. The linear approximation to the perturbation response is, from Eqs. (3) and (5)

$$\epsilon_n \approx \bar{K} \cdot \epsilon_\phi \quad (7)$$

The model parameter error vector is "proportional to" the Doppler phase error vector where the constant of proportionality is dependent on the given atmospheric model. From Eq. (7) it can be shown that the variances of the refractivity and Doppler phase vectors are linearly related by the factor K_{ij}^2 when the Doppler phase error components are assumed to be statistically independent:

$$D_n \approx \bar{K}^2 \cdot D_\phi \quad \bar{K}^2 = (K_{ij}^2) \quad (8)$$

where \bar{D}_n and \bar{D} are the variances of the refractivity and phase vector.

The K matrix represents the incremental response of the i th refractivity profile component to an incremental change in the j th Doppler phase component. The columns of the K matrix represent the directional derivatives of the inverse function σ^{-1} (assuming a unique inverse exists). From Eq. (4) and the definition of the directional derivative

$$\lim_{h \rightarrow 0} \frac{\sigma^{-1}(\phi + h\phi_i) - \sigma^{-1}(\phi)}{h}$$

may be used to evaluate the K matrix directly. When the refractivity of the planetary atmosphere is small, it may be shown that the operator σ is linear. In matrix form:

$$\phi = \bar{\sigma} \cdot n$$

Taking the inverse

$$n = \bar{\sigma}^{-1} \cdot \phi$$

from this form we see that

$$K = \bar{\sigma}^{-1}$$

If we examine the operator ϕ apply the definition of p and η , and discretize the integrand of Eq. (1) by assuming spherically stratified concentric shells, we obtain the summation

$$\Phi_j = \sum_k n \sec i_{jk} \Delta r \quad (9)$$

where i_{jk} = angle i at the k th shell and the j th ray, and Δr is the shell thickness, and $\sec i_{jk}$ is known for the j th ray path. Taking the inverse of the matrix form of Eq. (9) results in the matrix inverse operator ϕ^{-1} .

We next turn our attention to relating the variance vector of the received Doppler phase record to known quantities. We know the Doppler phase is the time integral of the Doppler frequency. The Doppler frequency shift is sampled every Δt sec, yielding

$$\Phi_j = \sum_{m=1}^I f_m \Delta t$$

where f_m = m th Doppler frequency sample. Assuming the Doppler frequency shift errors are statistically independent, the variance of the sum is equal to the sum of the variances.

$$\Gamma_{\Phi} = \begin{bmatrix} 1 & 0 & 0 & \dots \\ 1 & 1 & 0 & \dots \\ 1 & 1 & 1 & \dots \\ \dots & \dots & \dots & \dots \end{bmatrix} D_f \Delta t^2 \quad (10)$$

where D_f is the variance of the Doppler frequency shift. When we assume the mean frequency error is zero, the Doppler frequency shift variance is equal to the mean square Doppler frequency measurement noise:

$$D_{fi}^{1/2} = D_f^{1/2} = \text{rms frequency noise} \quad i = 1, 2, 3, \dots \quad (11)$$

Using Eqs. (8), (9), and (10), the standard deviation of the N unit refractivity divided by the $D_f^{1/2}$ [N unit = $(n-1) \times 10^6$] is plotted in figure 3 assuming 200 1-second samples. The $3\sigma(\pm 3\sigma)$ uncertainty region about a given model (Mars) is shown in figure 4 for an rms noise error of 0.10 Hz (approximate value for Mariner 4).

An uncertainty in the measured quantity space has been transformed to an uncertainty in the indirectly sensed quantity space by linearizing the response to small measurement errors. A region (volume) of confidence in the Doppler vector space is defined in the conventional manner and is assumed known. The corresponding region (volume) in the indirectly sensed quantity space is the confidence region of the model space. The uncertainty region has been plotted for several planetary atmospheres with a physically realistic assumption of the nature of the Doppler frequency error. This is intended to yield a quantitative appreciation for the deviation of the solution model caused by realistic values of random additive noise in the Doppler data.

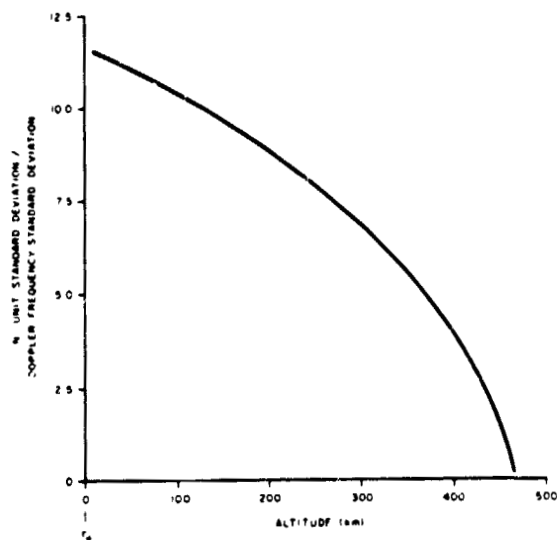


Figure 3.— N unit standard deviation for a unit Doppler standard deviation.

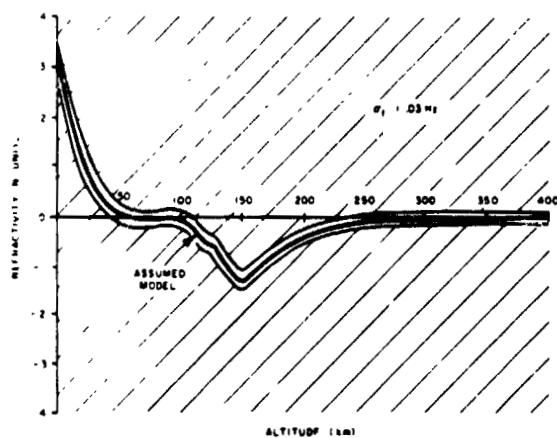


Figure 4.— 3σ uncertainty region.

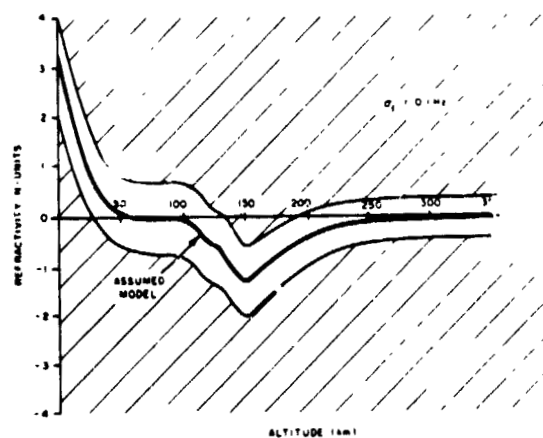


Figure 5.— 3σ uncertainty region.

AN OCCULTATION SATELLITE SYSTEM FOR DETERMINING PRESSURE LEVELS IN THE ATMOSPHERE

Steven G. Ungar and Bruce B. Lusignan, Stanford University

1 N73-11601

ABSTRACT

An operational two-satellite microwave occultation system will establish a pressure reference level to be used in fixing the temperature-pressure profile generated by the SIRS infrared sensor as a function of altitude. In the final error analysis, simulated data for the SIRS sensor were used to test the performance of the occultation system.

The results of this analysis indicate that the occultation system is capable of measuring the altitude of the 300-mb level to within about 24 mrms, given a maximum error of 2°K in the input temperature profile. The effects of water vapor can be corrected by suitable climatological profiles, and improvements in the accuracy of the SIRS instrument should yield additional improvements in the performance of the occultation system.

The final stage of the error analysis for the two-satellite microwave occultation pressure reference level system has been completed. In this stage, simulated data for the SIRS infrared (IR) sensor were used to test the performance of the occultation system. The effect of the oblateness of the earth was included in the analysis, as was the effect of water vapor. The results of this analysis are in agreement with previous results; that is, the occultation system is capable of measuring the altitude of the 300-mb level to within about 24 mrms, given a maximum error of 2°K in the temperature profile used as input for the system.

The simulated SIRS profiles were provided by Dr. William Smith of NOAA, and consist of 206 profiles of temperature versus pressure, with temperature given at 100 pressure levels between 1000-mb and 0.1-mb. A typical data set is shown in figure 1. Each level gives a simulated temperature "sensed" by the IR instrument, a temperature "error" (how this "sensed" temperature differs from the true temperature), and the error in height resulting from the accumulated error in temperature. In addition, each set tabulates rms errors for various ranges of pressure (e.g., 700 to 400-mb, 400 to 100-mb).

Half of the data sets give simulated profiles for a seven channel sensor, the other half for a nine channel sensor. Only data sets for the nine channel sensor were used in this error analysis. The data sets are divided into three groups labeled set 2, set 3, and set 4 according to the latitude of the stations in the set. These latitude ranges are indicated in table 1.

The pressure-reference level program was run as follows. The "sensed" temperature profile and the corresponding temperature "errors" were combined to give the "true" temperature profile. This profile was used as the "reference atmosphere."

TABLE 1.- LATITUDE RANGES
Set No.

Lat.	2	3	4
Min	25 N	38 N	51 N
Max	37 N	51 N	76 N

The original "sensed" temperature profile was used as "measured data." An atmosphere was constructed from this profile and shifted vertically until the results of ray tracing through the vertically shifted "measured" profile matched the results of ray tracing through the "reference atmosphere." In both the "reference atmosphere" and the "measured atmosphere" portions of the program, a random error of absolute magnitude 0.5 m was added to the phase defect to simulate the uncertainty in satellite spacing.

The resulting altitude error at each of nine pressure levels was recorded. A typical output sample is shown in figure 2. Statistical summaries for each of the three latitude ranges are given in tables 2 through 7. Tables 2 through 4 give the results for a satellite separation of 7863 km, while tables 5 through 7 give the results for a satellite separation of 7888 km. The greater the satellite separation, the lower the closest approach altitude of the ray. Table 8 lists the closest approach altitudes for each of the cases shown in table 2 through 7. The results are also shown in figures 3 and 4.

A smaller number of profiles were run with water vapor introduced into the reference atmosphere portion of the program. The water vapor information consisted of the associated rawinsonde relative humidity profiles. A simplistic

STATION NO. 72201 PROFILE NO. 1 CASE NO. 2 SET NO. 2

KEY WEST/INT FLORIDA

24 33 N 81 45 W RADIUS OF EARTH AT THIS LATITUDE 6374.4727 KM.

LEVEL (MB)

500.000	-1.32
400.000	-1.09
300.000	0.50
200.000	-25.81
140.000	-55.26
100.000	-45.30
70.000	-18.83
50.000	3.07
30.000	-13.75

STATISTICS FOR LOWER 9 PRESSURE LEVELS

MEAN	-17.50
ST. DEV.	21.15
RMS	27.45
MIN	-55.26
MAX	3.07

Figure 2.- Typical output of pressure reference system.

TABLE 2.- STATISTICS FOR 14 STATIONS
SATELLITE SEPARATION 7863.0000 km - SET 2

Level (mb)	Mean	SD	RMS	Min	Max
500.000	7.70	14.05	16.02	-16.89	28.33
400.000	8.18	16.42	18.35	-21.21	30.48
300.000	17.64	28.40	33.43	-30.50	53.93
200.000	24.65	35.57	43.28	-25.81	92.03
140.000	2.64	33.55	33.65	-55.26	61.61
100.000	0.57	27.10	27.10	-45.36	58.87
70.000	4.56	28.17	28.54	-42.50	56.64
50.000	11.02	38.20	39.76	-58.88	91.92
30.000	10.55	37.38	38.84	-49.27	58.25

TABLE 3.- STATISTICS FOR 20 STATIONS
SATELLITE SEPARATION 7863.0000 km - SET 3

Level (mb)	Mean	SD	RMS	Min	Max
500.000	1.21	17.67	17.71	-20.61	36.88
400.000	-3.15	12.97	13.35	-24.05	17.89
300.000	-0.07	16.47	16.47	-43.97	21.99
200.000	6.28	25.77	26.52	-51.67	41.15
140.000	4.90	42.22	42.50	-83.59	80.43
100.000	-2.32	39.53	39.59	-73.74	95.77
70.000	1.13	28.35	28.38	-33.48	87.17
50.000	9.72	24.72	26.57	-28.99	77.82
30.000	24.42	40.21	47.05	-47.99	87.49

TABLE 4.- STATISTICS FOR 18 STATIONS
SATELLITE SEPARATION 7863.0000 km - SET 4

Level (mb)	Mean	SD	RMS	Min	Max
500.000	12.38	22.83	25.97	-27.09	48.48
400.000	9.56	25.29	27.04	-31.75	42.22
300.000	-0.82	25.32	25.33	-49.44	38.81
200.000	12.11	22.26	25.34	-12.92	72.49
140.000	14.80	29.62	33.11	-21.11	95.06
100.000	15.39	33.36	36.74	-38.29	105.33
70.000	11.87	29.70	31.99	-41.17	76.05
50.000	1.35	19.38	19.42	-31.85	30.24
30.000	-21.75	36.20	42.23	-78.40	43.18

TABLE 5.- STATISTICS FOR 14 STATIONS
SATELLITE SEPARATION 7888.0000 km - SET 2

Level (mb)	Mean	SD	RMS	Min	Max
500.000	2.72	16.82	17.04	-32.38	30.76
400.000	3.20	17.42	17.71	-28.07	28.50
300.000	12.66	15.84	20.27	-14.57	34.83
200.000	19.67	22.48	29.87	-13.67	59.76
140.000	-2.34	32.35	32.44	-47.95	48.25
100.000	-4.41	28.62	28.96	-62.93	42.37
70.000	-0.42	23.34	23.34	-47.52	32.24
50.000	6.04	29.24	29.86	-32.20	68.65
30.000	5.57	25.40	26.00	-44.91	44.65

TABLE 6.- STATISTICS FOR 19 STATIONS
SATELLITE SEPARATION 7888.0000 km - SET 3

Level (mb)	Mean	SD	RMS	Min	Max
500.000	-3.31	19.62	19.90	-43.18	27.33
400.000	-7.95	28.79	29.87	-68.50	41.94
300.000	-5.63	45.71	46.06	-96.07	69.94
200.000	-0.87	43.21	43.22	-80.68	71.70
140.000	-0.31	38.60	38.60	-68.58	68.34
100.000	-6.19	37.56	38.12	-60.89	75.51
70.000	-2.94	30.74	30.88	-51.80	72.04
50.000	4.29	25.10	25.47	-33.07	51.54
30.000	15.71	37.80	40.93	-47.12	89.89

TABLE 7.- STATISTICS FOR 18 STATIONS
SATELLITE SEPARATION 7888.0000 km - SET 4

Level (mb)	Mean	SD	RMS	Min	Max
500.000	14.07	35.69	38.36	-29.69	90.63
400.000	11.25	28.34	30.49	-17.20	76.93
300.000	0.87	27.49	27.51	-40.40	56.26
200.000	13.81	49.56	51.45	-56.19	103.93
140.000	16.49	58.22	60.51	-60.69	133.87
100.000	17.08	54.47	57.08	-40.43	140.75
70.000	13.56	50.62	52.40	-47.56	127.72
50.000	3.04	49.95	50.04	-79.01	109.18
30.000	-20.06	63.61	66.70	-134.16	83.86

TABLE 8.- DRY ATMOSPHERE

Sat				Sep		
7863				7888		
Set	No. Prof	Min. alt.	Max. alt.	No. Prof	Min. alt.	Max. alt.
2	14	5.78	7.86	14	3.59	5.44
3	20	7.55	9.75	19	4.85	6.81
4	18	9.59	12.56	18	7.29	9.91

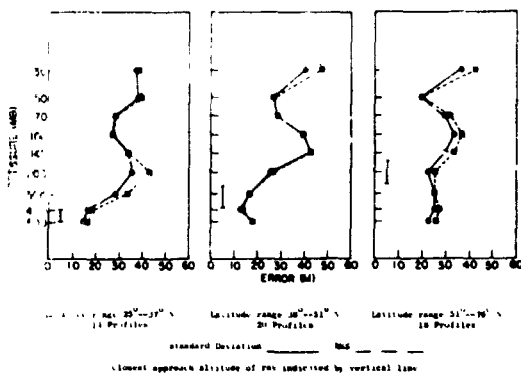


Figure 3.— Standard deviation and rms error versus pressure. Dry atmosphere satellite separation 7863.0 km.

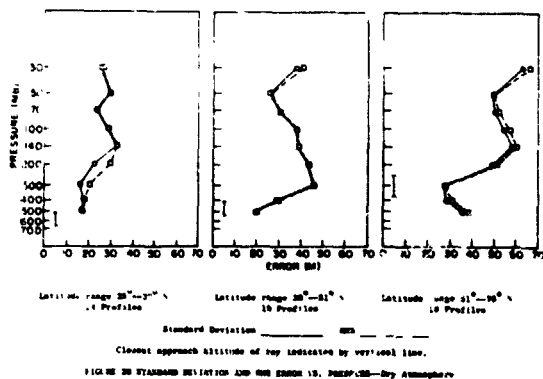
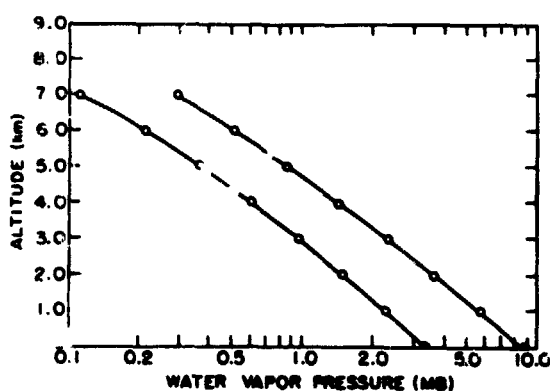


Figure 4.— Standard deviation and rms error versus pressure. Dry atmosphere satellite separation 7888.0 km.



Profiles were adapted from Figure 6, Hutcherson, D.C., "Water Vapor in the Atmosphere," Humidity and Moisture, Vol. 2, 1965, pp. 486-494.

The above profiles were extrapolated to 25 km by use of the equation $e(z) = e(7) \exp[0.5476 \cdot (z-7)]$, where e is the water vapor pressure, in millibars, and z is the altitude, in kilometers.

Figure 6.— Water vapor profiles.

water vapor correction was used in the second half ("measured data" portion) of the program. This simplistic correction consisted in each case of two humidity profiles, representing either a "high" value of water vapor or a "low" value (see fig. 5). The appropriate corrective profile was chosen based on the surface value of the rawinsonde water vapor profile. All the data sets from sets 3 and 4, and one of the profiles from set 2, used the lower corrective profile; the remaining profiles in set 2 used the higher corrective profile. The results of this run are summarized in tables 9 through 14. Tables 9 through 11 give the results for a satellite separation of 7863 km, and tables 12 through 14 give the results for a satellite separation of 7888 km. The corresponding closest approach altitudes are listed in table 15. The results are shown in figures 6 and 7.

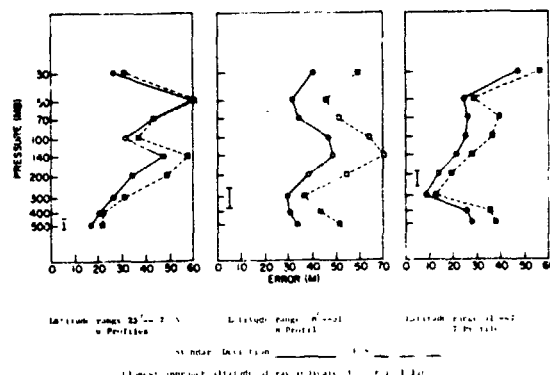


Figure 6.— Standard deviation and rms error versus pressure. Wet atmosphere with simplistic climatic correction satellite separation 7863.0 km.

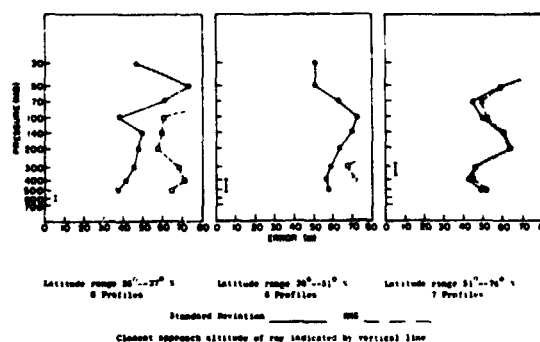


Figure 7.— Standard deviation and rms error versus pressure. Wet atmosphere with simplistic climatic correction satellite separation 7888.0 km.

TABLE 9. - STATISTICS FOR 6 STATIONS
SATELLITE SEPARATION 7863.0000 km - SET 2

Level (mb)	Mean	SD	RMS	Min	Max
500.000	13.12	16.80	21.31	-7.95	40.43
400.000	8.31	20.71	21.85	-23.56	29.32
300.000	15.22	26.48	30.54	-27.64	40.54
200.000	34.51	34.32	48.67	-2.97	93.47
140.000	32.68	47.46	57.63	-18.32	80.53
100.000	19.16	31.17	36.59	-26.20	59.18
70.000	6.13	43.39	43.82	-56.23	46.92
50.000	11.06	59.10	60.13	-52.70	83.33
30.000	15.58	26.30	30.57	-19.93	46.52

TABLE 10. - STATISTICS FOR 8 STATIONS
SATELLITE SEPARATION 7863.0000 km - SET 3

Level (mb)	Mean	SD	RMS	Min	Max
500.000	39.18	33.84	51.77	-5.46	99.36
400.000	30.87	30.54	43.42	-1.78	88.95
300.000	21.53	29.34	36.39	-16.31	68.32
200.000	38.97	38.62	54.86	-3.64	94.15
140.000	50.94	48.82	70.56	-22.24	121.11
100.000	44.02	46.73	64.20	-35.13	105.45
70.000	38.30	34.71	51.69	-13.50	88.64
50.000	32.68	31.24	45.21	-21.36	70.95
30.000	43.89	40.11	59.45	-33.54	85.94

TABLE 11. - STATISTICS FOR 7 STATIONS
SATELLITE SEPARATION 7863.0000 km - SET 4

Level (mb)	Mean	SD	RMS	Min	Max
500.000	25.19	27.94	37.62	-20.77	66.63
400.000	23.62	25.98	35.12	-26.80	52.06
300.000	9.41	8.46	12.66	-3.93	18.65
200.000	13.59	14.00	19.51	-7.75	35.83
140.000	17.46	21.65	27.81	-13.57	41.90
100.000	25.93	25.36	36.27	-23.61	56.55
70.000	29.48	26.43	39.59	-20.16	69.11
50.000	14.41	24.78	28.66	-19.81	43.91
30.000	-30.30	47.90	56.68	-91.29	39.23

TABLE 12. - STATISTICS FOR 6 STATIONS
SATELLITE SEPARATION 7888.0000 km - SET 2

Level (mb)	Mean	SD	RMS	Min	Max
500.000	-53.11	37.21	64.85	-97.58	8.24
400.000	-57.93	41.29	71.14	-122.42	-2.86
300.000	-51.02	45.27	68.21	-126.50	0.03
200.000	-31.72	47.98	57.52	-68.77	61.28
140.000	-33.55	49.06	59.44	-89.27	48.35
100.000	-47.07	37.68	60.29	-77.58	27.00
70.000	-60.11	60.83	85.52	-142.15	9.67
50.000	-55.17	73.51	91.91	-151.57	12.85
30.000	-50.65	46.45	68.73	-118.79	5.25

TABLE 13. - STATISTICS FOR 8 STATIONS
SATELLITE SEPARATION 7888.0000 km - SET 3

Level (mb)	Mean	SD	RMS	Min	Max
500.000	51.29	57.44	77.00	-12.32	146.25
400.000	42.97	56.47	70.96	-31.52	135.84
300.000	33.63	58.84	67.77	-51.88	115.21
200.000	51.08	63.97	81.86	-32.43	135.99
140.000	63.04	69.96	94.18	-30.39	168.00
100.000	56.13	72.17	91.42	-43.29	150.06
70.000	50.40	63.12	80.77	-21.67	126.30
50.000	50.12	50.68	71.27	-2.69	117.84
30.000	55.99	50.55	75.44	-1.54	125.75

TABLE 14. - STATISTICS FOR 7 STATIONS
SATELLITE SEPARATION 7888.0000 km - SET 4

Level (mb)	Mean	SD	RMS	Min	Max
500.000	15.78	48.91	51.40	-34.20	99.85
400.000	14.21	42.20	44.53	-37.26	85.28
300.000	0.01	46.03	46.03	-67.46	67.63
200.000	4.18	64.08	64.22	-89.11	92.39
140.000	8.06	60.09	60.63	-94.93	75.12
100.000	16.53	49.55	52.23	-68.07	67.30
70.000	20.07	44.89	49.18	-41.15	72.56
50.000	5.00	59.20	59.41	-65.16	75.32
30.000	-39.71	93.34	101.44	-145.74	94.39

TABLE 15. - WET ATMOSPHERE WITH SIMPLISTIC
CLIMATIC CORRECTION

		Sat		Sep			
		7863		7888			
Set	No. Prof	Min. alt.	Max. alt.	No. Prof	Min. alt.	Max. alt.	
2	6	0.25	7.59	6	4.11	4.95	
3	8	7.02	9.96	8	4.99	7.10	
4	7	9.70	11.47	7	7.40	9.28	

To determine the effect of an improvement in the water vapor correction technique—that is, the use of more sophisticated techniques than the simplistic method described above—the same data were run with the exact rawinsonde water vapor profile used in both the reference atmosphere and in the “measured” atmosphere. This is equivalent to an ideal water vapor correction. Results of this run (for the 7863-km satellite separation only) are given in tables 16 through 18, and in figure 8. The corresponding closest approach altitudes are shown in table 19.

Finally, to determine the effects of the temperature errors on the system, the analysis of six low altitude profiles (without water) was repeated, this time using the reference temperature values (the values without any error) in both the reference atmosphere and in the “measured” atmosphere. This is equivalent to having an ideal temperature sensor. The result of this run is shown in figure 9, for the 7863-km satellite separation only.

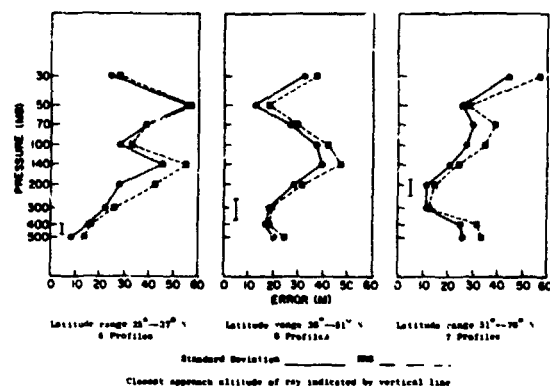


Figure 8.— Standard deviation and rms error versus pressure. Wet atmosphere with ideal water vapor correction satellite separation 7863.0 km.

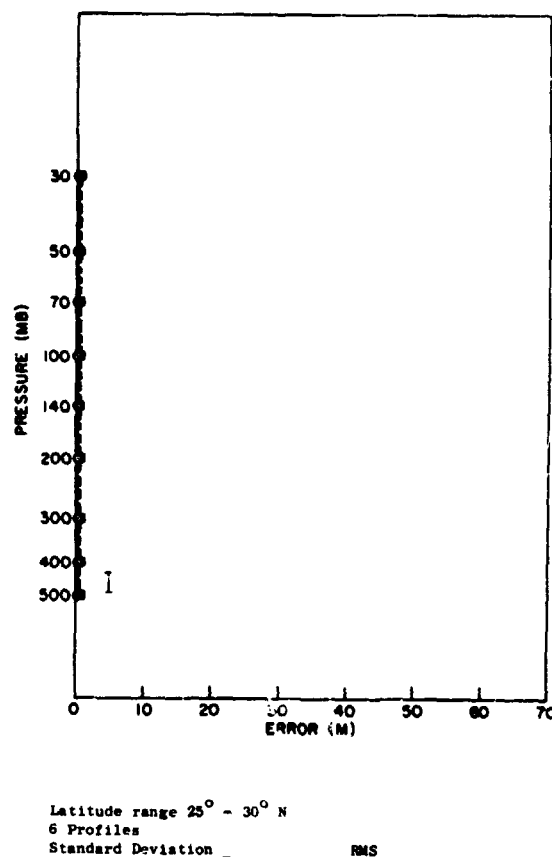


Figure 9.— Standard deviation and rms error versus pressure. Ideal temperature profile, dry atmosphere.

ANALYSIS

In figures 3, 4, 6, and 7, the solid line represents the standard deviation of the error (in meters) while the stippled line indicates the rms error; both are shown as a function of pressure. The difference between the standard deviation and the rms values at each pressure level is an indication of the size of the mean error at this level. The rms value is equal to the square root of the sum of the squares of the standard deviation and the mean ($\text{rms}^2 = \text{SD}^2 + \text{mean}^2$).

A cursory examination of these figures shows that the curve of altitude error versus pressure follows a typical pattern, regardless of whether the atmosphere is wet or dry, or, if wet, how the water vapor error is corrected.

TABLE 16.- STATISTICS FOR 6 STATIONS
SATELLITE SEPARATION 7863.0900 km - SET 2

Level (mb)	Mean	SD	RMS	Min	Max
500.000	10.37	8.66	13.51	-0.26	21.08
400.000	5.56	15.49	16.45	-19.55	28.87
300.000	12.47	22.52	25.74	-23.63	45.27
200.000	31.76	27.86	42.25	1.24	71.97
140.000	29.93	45.59	54.54	-23.13	80.99
100.000	16.42	27.90	32.38	-18.49	46.27
70.000	3.38	38.68	38.83	-54.34	56.15
50.000	8.31	55.82	56.43	-49.15	92.56
30.000	12.83	24.88	28.00	-15.91	55.76

TABLE 17.- STATISTICS FOR 8 STATIONS
SATELLITE SEPARATION 7863.0000 km - SET 2

Level (mb)	Mean	SD	RMS	Min	Max
500.000	13.91	20.20	24.60	-18.24	38.74
400.000	5.59	17.51	18.38	-19.58	27.90
300.000	-3.74	18.36	18.74	-39.93	24.18
200.000	13.70	28.19	31.34	-20.49	66.59
140.000	25.67	39.40	47.02	-40.89	92.33
100.000	18.75	37.49	41.92	-53.79	77.89
70.000	13.02	26.43	29.47	-32.17	61.07
50.000	12.74	12.80	18.06	2.70	42.34
30.000	18.61	32.47	37.43	-46.31	67.28

TABLE 18.- STATISTICS FOR 7 STATIONS
SATELLITE SEPARATION 7863.0000 km - SET 4

Level (mb)	Mean	SD	RMS	Min	Max
500.000	20.38	26.00	33.03	-25.66	48.63
400.000	18.81	25.17	31.42	-31.70	45.03
300.000	4.60	11.40	12.29	-9.59	19.66
200.000	8.78	11.17	14.20	-8.24	16.14
140.000	12.65	20.65	24.22	-13.66	34.70
100.000	21.12	27.21	34.45	-28.50	58.59
70.000	24.67	29.83	38.71	-25.05	71.14
50.000	9.60	25.67	27.41	-20.39	45.94
30.000	-35.12	44.69	56.83	-100.97	18.59

TABLE 19.- WET ATMOSPHERE WITH
IDEAL CORRECTION

Set	No. Prof	Sat	Sep
		Min. alt.	Max. alt.
2	6	6.11	7.65
3	8	7.69	10.00
4	7	9.98	11.85

Typically there are two minima and two maxima. One of the minima generally occurs in the neighborhood of 300 to 400 mb, and the second in the neighborhood of 50 to 70 mb. The maxima are near the 30-mb level and between 100 and 200 mb (which, for most profiles, is the pressure level of the tropopause).

The pattern may be explained in the following way: the minimum at the lower altitude, near the 300-mb level, is associated with the closest approach altitude of the ray path (indicated by the vertical line in each figure). The system is most sensitive to atmospheric effects in approximately the lower 3 km of the ray path, so the absolute value of error in this neighborhood tends to be minimized.

The overall pattern is due to the errors in the temperature profile, which generate errors in altitude when the temperature-altitude profile is constructed using the hypsometric formula. This effect is illustrated by the absence of any extrema at all in the results of the ideal temperature profile case (fig. 9). Here each level has the same error as that of the 300- or 400-mb level.

The magnitude of the error is, of course, of primary interest. We see in figures 3 and 4 that the rms and standard deviation of the error below 300 mb are generally around 20 to 30 m. Figure 10 shows the 300-mb error for each of the profiles used in the generation of figure 3, plotted as a function of rms temperature error in the 100- to 400-mb pressure range. While there is no clear regression line, it should be noted that the upper left-hand portion of the chart is free of any data points. This indicates that a low rms value of temperature error in the 100- to 400-mb pressure range invariably results in a small altitude error at the 300-mb level. It is therefore reasonable to expect that improvements in the performance of the SIRS instrument will lead to concomitant improvements in the performance of the occultation system. (It should also be noted that a high rms value of temperature error does not necessarily lead to a large error in the altitude of the 300-mb level, as witnessed by the large number of points in the lower right-hand portion of the chart.)

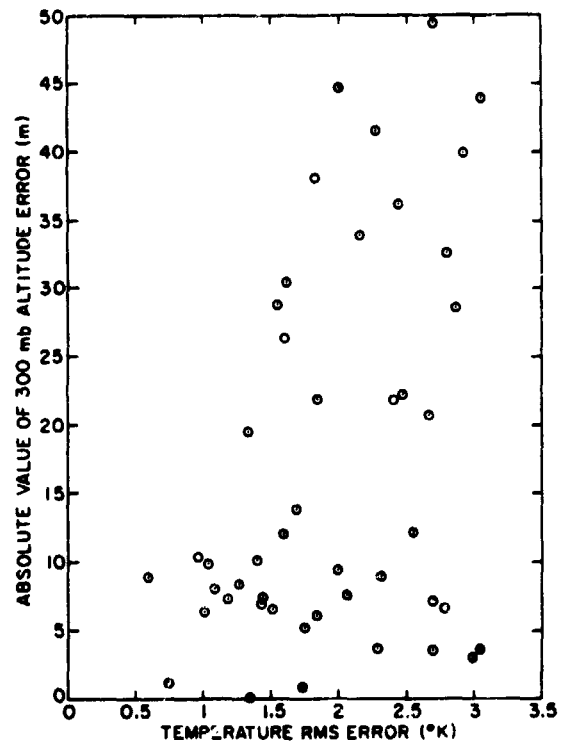


Figure 10.— Altitude error at 300-mb versus temperature rms error (pressure range 100-400 mb).

Comparison of figures 3 and 4 shows that it is possible to choose a satellite separation for each of the latitude ranges such that the rms and standard deviation at the 300-mb level is less than 26 m. We do not suggest that there must necessarily be more than one satellite pair in the system, but merely note that minimum ray altitude is an important parameter that should be chosen in such a way as to optimize the system performance. In each of the figures, the lower minimum occurs in the range of the closest approach altitude.

Turning to figure 6, we see that the simplistic water vapor correction gives remarkably good results. RMS and standard deviation at the 300-mb level and below are once again in the neighborhood of 20 to 30 m. That the errors are due to the failure of the corrective profile to adequately account for the water vapor is illustrated by figure 8, which gives the results for the "ideal" water vapor correction. The results shown in figure 8 should be compared to the corresponding curves of figure 3. It is apparent that errors introduced by water vapor can be eliminated by properly correcting for the water vapor. If exact water vapor corrections are not available, figure 6 shows that even the most simplistic of corrections, in which a standard water vapor profile is used to correct for water vapor effects, can have dramatic results. The effect of not correcting for water vapor at all is to obtain errors on the order of hundreds of meters.

The effect of varying minimum ray height on the system performance in the presence of water vapor is illustrated in figure 7. We see that the system operating at the lower ray height (greater satellite separation) yields generally poorer results than the system operating at the greater ray height (smaller satellite separation). Since water vapor concentration falls off rapidly with altitude (scale height on the order of 2 km), this is not a particularly surprising result. Proper water vapor correction is more important as the ray passes into regions of high water vapor content. Although it was not possible to run the experiment of figure 8 for the lower ray height, presumably an "ideal" water vapor correction would have the same dramatic result in this case as it had at the higher ray altitude.

CONCLUSIONS

Based on the analysis of the simulated SIRS data, we can say with confidence that the microwave occultation system is capable of establishing a 300-mb pressure reference level to within about 24 m. The effects of water vapor can be corrected by suitable climatological profiles. Our work has shown that even the most simplistic climatological approach results in a dramatic improvement in accuracy; more sophisticated methods are bound to be even more effective.

Improvements in the accuracy of the SIRS instrument will yield improvements in the performance of the occultation system. Most of the errors in the output of the system (neglecting those due to water vapor) are due to error in the input temperature profile. As these input errors are reduced, system performance will be significantly improved. But even with the present SIRS instrumentation, the microwave occultation system gives results that are acceptable to, and badly needed by, the meteorological community.

DISCUSSION

Gille: Does this give the absolute altitude of the pressure reference above the surface?

Lusignan: Yes. It is referred to as the geopotential of the earth.

Gille: How sensitive is that to orbit variations?

Lusignan: The system relies on knowing differential movement between the satellite, to the order of 0.5 m. It is similar to the transponder used in Venus and Mars experiments.

Reinisch: You are making phase measurements; at 6-cm wavelength, how great is the uncertainty of the separation and how do you take care of modulo 2π phase change?

Lusignan: There are two systems, parallel to those of the Mariner flights. There is a phase system, a coherent transponder, which gives an accuracy of 3-cm. This is a differential measurement. The second system has an FM modulation superimposed that gives an absolute reference to the order of 0.5 m.

Wright: Is there an ionospheric version of this experiment under consideration?

Lusignan: Yes. Dr. Grossi has proposed dual satellite ionosphere studies.

Rodgers: What effect will horizontal inhomogeneities in water vapor and temperature have on your measurements?

Lusignan: We've run tests with horizontal profiles in them. Profiles across the path have negligible effect. Along the path we are sensitive to an integrated average over about 300 to 500 km. We measure the weighted average of the pressure in that distance.

Reinisch: Do you expect a lot of phase scintillation that would distort your measurement?

Lusignan: In the lower part of the atmosphere where there is water vapor, we do get some phase scintillation, but it does not affect our group path measurement. We do expect some clean multipath in the lower part when in the range 500-700 mb. Normally we operate at an altitude above where this occurs (above 300 mb).

OCCULTATION MEASUREMENTS

BIBLIOGRAPHY

- Backus, G., and Gilbert, J. F., Numerical Applications of a Formalism for Geophysical Inverse Problems, *Geophys. J. Roy. Astron. Soc.*, **13**, 247-276, 1967.
- Backus, G., and Gilbert, J. F., The Resolving Power of Gross Earth Data, *Geophys. J. Roy. Astron. Soc.*, **16**, 169-205, 1968.
- Backus, G., and Gilbert, J. F., Uniqueness in the Inversion of Inaccurate Gross Earth Data, *Phil. Trans. Roy. Soc. London*, **266**, 123-192, March 5, 1970.
- Baum, W. A., and Code, A. D., A Photometric Observation of the Occultation of Sigma Arietis by Jupiter, *Astron. J.*, **58**, 108-112, 1953.
- Beers, Y., Introduction to the Theory of Error, Addison-Wesley, Reading, Mass., 1957.
- Bennett, J. A., The Calculation of Doppler Shifts Due to a Changing Ionosphere, *J. Atmos. Terr. Phys.*, **29**, 887-891, 1967.
- Bennett, J. A., The Ray Theory of Doppler Frequency Shifts, *Aust. J. Phys.*, **21**, 259-272, 1968.
- Bennett, J. A., On the Application of Variation Techniques to the Ray Theory of Radio Propagation, *Radio Sci.*, **4**, 667-678, Aug. 1969.
- Cain, D. L., Kliore, A. J., and Levy, G. S., The Mariner 4 Occultation Experiment: Summary of Data Reduction Methods, AIAA Paper No. 66-148, AIAA 3rd Aerospace Sciences Meeting, New York, New York, January 24-26, 1966.
- Chamberlain, J. W., and McElroy, M. B., The Martian Atmosphere: An Interpretation of the Mariner Occultation Experiment, *Science*, **152**, 21, 1966.
- Coogan, J. M., A Method for Studying Planetary Atmospheres Employing the Dual Flyby Mode, AIAA 5th Aerospace Conference, New York, New York, Jan. 23-26, 1967.
- de Vaucouleurs, G., A Survey of Physical Problems of the Nearer Planets and a Review of Observational Techniques Applicable to Balloon-Borne Telescope Systems, Ch. 2 in Strong J. D., de Vaucouleurs G., and Zwicky P., "Planetary Astronomy from Satellite-Substitute Vehicles", Missile Development Center, Holloman Air Force Base, New Mexico, Report TR-60-6, March 1960.
- Elsmore, B., Radio Observations of the Lunar Ionosphere, *Phil. Mag.*, **2**, 1040-1046, 1957.
- Eshleman, V. R., Solar System Radio Astronomy, Proceedings of the Advanced Study Institute on Solar System Radio Astronomy, Sounion, Greece, 1964, J. Aarons, ed., Plenum Press, New York, Ch. 14, 267-293, 1965.
- Eshleman, V. R., Radar Astronomy, *Science*, **158**, 585-597, Nov. 3, 1967.
- Eshleman, V. R., Fjeldbo, G., Anderson, J. D., Kliore, A., Dyce, R. B., Venus: Lower Atmosphere Not Measured, *Science*, **162**, 661-662, 8 Nov. 1968.

Esposito, R., and Schumer, M. A., Probing Techniques for Transmission Absorption Spectroscopy of Planetary Atmospheres Using Satellites, Remote Sensing of Environment, 1, 81-91, 1969/1970

Fabry, C., Le Role des Atmospheres dan les Occultations par les Planetes, J. Observateurs, 12, 1-10, 1929.

Fischbach, F. F., Remarks on the Inference of Neutral Atmospheric Parameters from Microwave Occultation Data, JPL Tech. Report 32-1475, 99-103, July 1, 1970.

Fischbach, F. F., Graves, M. E., Hays, P. B., Klein, G. S., Mizgala, C. M., and Peterson, J. W., Atmospheric Sounding by Satellite Measurements of Stellar Refraction, Univ. of Michigan ORA Tech. Rep. 04963-2-T, Dec. 1962.

Fischbach, F. F., Graves, M. E., Hays, P. B., and Roble, R. G., Satellite Measurement of Atmospheric Structure by Stellar Refraction, Univ. Michigan ORA Tech. Rep. 06647-1-T, April 1965.

Fischbach, F. F., Graves, M. E., and Jones, L. M., Satellite Meteorology by Microwave Refraction: Some Problems, COSPAR XIth Plenary Meeting, Tokyo, Space Research IX, North Holland, 1969.

Fjeldbo, G., Bistatic-Radar Methods for Studying Planetary Ionospheres and Surfaces, Stanford Univ., Rep. SU-SEL-64-025, Stanford Electronics Lab, Stanford University, 1964a.

Fjeldbo, G., Brief Summary of Methods Used to Study Model Atmospheres for Mars, Stanford Univ. memo. dated 1964 b.

Fjeldbo, G., Radio Occultation Measurements of Planetary Atmospheres and Planetary Surface Topography, AIAA 5th Aerospace Sciences Meeting, Jan. 23-26, 1967.

Fjeldbo, G., and Eshleman, V. R., The Bistatic Radar Occultation Method for the Study of Planetary Atmospheres, J. Geophys. Res., 70, 3217-3226, 1965.

Fjeldbo, G., and Eshleman, V. R., The Atmosphere of Mars Analysed by Integral Inversion of the Mariner IV Occultation Data, Stanford Univ. Rep. No. SU-SEL-67-109, Nov. 1967.

Fjeldbo, G., and Eshleman, V. R., The Atmosphere of Mars Analyzed by Integral Inversion of the Mariner IV Occultation Data, Planet. Space Sci., 16, 1035-1059, 1968.

Fjeldbo, G., and Eshleman, V. R., The Atmosphere of Mars: Radio Occultation Measurements and Interpretations, The Atmospheres of Venus and Mars, J. C. Brandt and M. B. McElroy, eds., Gordon and Breach, N.Y., 1968.

Fjeldbo, G., and Eshleman, V. R., The Atmosphere of Venus as Studied with the Mariner 5 Dual Radio Frequency Occultation Experiment, Stanford Univ. Rep. SU-SEL-69-003, Jan. 1969.

Fjeldbo, G., and Eshleman, V. R., The Atmosphere of Venus as Studied with the Mariner 5 Dual Radio Frequency Occultation Experiment, Radio Sci., 4, 879-897, 1969.

- Fjeldbo, G., Eshleman, V. R., Garriott, O. K., and Smith, F. L., The Two-Frequency Bistatic Radar-Occultation Method for the Study of Planetary Atmospheres, J. Geophys. Res., **70**, 3701-3710, 1965.
- Fjeldbo, G., Eshleman, V. R., Kliore, A. J., Cain, D. L., Levy, G. S., and Drake, F. D., Preliminary Results of the Mariner IV Radio Occultation Measurements of the Upper Atmosphere of Mars, Proc. of Caltech-JPL Lunar and Planetary Conf., 267-272, Sept. 13-18, 1965.
- Fjeldbo, G., Fjeldbo, W. C., and Eshleman, V. R., Models for Atmosphere of Mars Based on the Mariner 4 Occultation Experiment, J. Geophys. Res., **71**, 2307-2316, May 1, 1966 a.
- Fjeldbo, G., Fjeldbo, W. C., and Eshleman, V. R., Atmosphere of Mars: Mariner IV Models Compared, Science, **151**, 1518-1523, 1966 b.
- Fjeldbo, G., Kliore, A., and Eshleman, V. R., The Neutral Atmosphere of Venus as Studied with the Mariner 5 Radio Occultation Experiments, Astronom. J., **76**, 123-140, March 1971.
- Fjeldbo, G., Kliore, A. J., and Seidel, B., The Mariner 1969 Occultation Measurements of the Upper Atmosphere of Mars, Radio Sci., **5**, 381-386, 1970.
- Gerver, M. L., and Markushевич, V. M., Determination of the Seismic Wave Velocity from the Travel-Time Curve, Geophys. J. Roy. Astron. Soc., **11**, 165-173, 1966.
- Gerver, M. L., and Markushевич, V. M., On the Characteristic Properties of Travel-Time Curves, Geophys. J. Roy. Astron. Soc., **11**, 241-246, 1967.
- Goldsmith, D. W., Differential Refraction in Planetary Atmospheres with Linear Scale Height Gradients, Icarus, **2**, 341-349, 1963.
- Graves, M. E., and Fischbach, F. F., Analysis of Microwave Occultation Techniques for Atmospheric Soundings, Tech. Rep. 05863-16-T, High Altitude Engineering Lab., U. Michigan, 1969.
- Grossi, M. D., Lee, J. F., Shear, I., Bravoco, R. R., Langevin, P. E., Goff, R. W., Boeitz, M., and Stiffler, J. J., Electromagnetic Probing of the Mars and Venus Atmospheres and Ionospheres from an Orbiting Pair, Raytheon Rep. R69-4374-1, Sudbury, Mass., Final Report on Contract NASw-1772, Sept. 19, 1969.
- Hagfors, T., Reflection and Diffraction in a Bistatic Radar Occultation Experiment, MIT Lincoln Lab Rep., August 2, 1967.
- Hamel, G., Integralgleichungen: Einführung in Lehre und Gebrauch, Springer Verlag, Berlin, 1937.
- Harrington, J. V., Study of a Small Solar Probe (Sunblazer) Part I, Radio Propagation Experiment, MIT Rep. PR-5255-5, July 1, 1965.
- Harrington, J. V., Goff, R. J., Grossi, M. D., Langworthy, B. M., Electromagnetic Measurements of Planetary Atmospheres and Ionospheres by an Orbital Pair, Raytheon IRP Rep. FR-66-71, March 7, 1966.

Harrington, J. V., and Grossi, M. D., Global Probing of the Earth's Atmosphere and Ionosphere by a Satellite-to-Satellite Radio Occultation Method, 3rd National Conf. on Aerospace Meteorology, New Orleans, La., May 6-9, 1968.

Harrington, J. V., Grossi, M. D., Goff, R. W., and Langworthy, B. M., Radio Occultation Measurements of Planetary Atmospheres and Ionospheres from an Orbiting Pair, AIAA Paper No. 69-53, AIAA 7th Aerospace Sciences Meeting, New York, New York, January 20-22, 1969.

Harrington, J. V., Grossi, M. D., and Langworthy, B. M., Mars Mariner IV Radio Occultation Experiment: Comments on the Uniqueness of the Results, J. Geophys. Res., 73, 3039-3041, 1968.

Hays, P. B., and Fischbach, F. F., Analytic Solution for Atmospheric Density from Satellite Measurements of Stellar Refraction, U. Michigan ORA Tech. Rep. 04963-3-T, Jan. 1963.

Hays, P. B., and Roble, R. G., Atmospheric Properties from the Inversion of Planetary Occultation Data, Planet. Space Sci., 16, 1197-1198, 1968.

Herglotz, G., Über das Benndorfsche Problem der Fortpflanzungsgeschwindigkeit der Erdenstrahlen, Phys. Z., 8, 145, 1907.

Hunten, D. M., and McElroy, M. B., The Upper Atmosphere of Venus: The Regulus Occultation Reconsidered, J. Geophys. Res., 73, 4446-4448, 1968.

Jones, L. M., Fischbach, F. F., and Peterson, J. W., Atmospheric Measurements from Satellite Observations of Stellar Refraction, U. Michigan ORA Tech. Rep. 04963-1-T, Jan. 1962.

Jones, L. M., Fischbach, F. F., and Peterson, J. W., Satellite Measurements of Atmospheric Structure by Refraction, Planet. Space Sci., 9, 351-352, June 1962.

Kliore, A. J., and Cain, D. L., Mariner 5 and the Radius of Venus, J. Atmos. Sci., 25, 549-554, 1968.

Kliore, A. J., Cain, D. L., Drake, F. D., Eshleman, V. R., Fjeldbo, G., and Levy, G. S., A Summary of Preliminary Results of Mariner IV Radio Occultation Experiment, Proc. Ionospheric Research Committee of the Avionics Panel, 75-81, AGARD/NATO, Rome, Italy, Sept. 21-25, 1965.

Kliore, A. J., Cain, D. L., and Hamilton, T. W., Determination of Some Physical Properties of the Atmosphere of Mars from Changes in the Doppler Signal of the Spacecraft on an Earth-Occultation Trajectory, JPL Tech. Rep. 32-674, Oct. 15, 1964.

Kliore, A. J., Cain, D. L., and Levy, G. S., Radio Occultation Measurements of the Martian Atmosphere Over Two Regions by the Mariner IV Space Probe, Moon and Planets, 226-239, North-Holland Publishing Co., 1966.

Kliore, A. J., Cain, D. L., and Levy, G. S., Radio Occultation Measurement of the Martian Atmosphere Over Two Regions by the Mariner 4 Space Probe, Space Research, VII, Moon and Planets, North Holland Publishing Company, Amsterdam, 1968.

Kliore, A. J., Cain, D. L., Levy, G. S., Eshleman, V. R., Drake, F. D., and Fjeldbo, G., The Mariner 4 Occultation Experiment, Astron. Aeron., 7 72-80, 1965.

Kliore, A. J., Cain, D. L., Levy, G. S., Eshleman, V. R., Fjeldbo, G., and Drake, F. D., Occultation Experiment: Results of the First Direct Measurement of Mars Atmosphere and Ionosphere, Science, 149, 1243-1248, Sept. 10, 1965.

Kliore, A. J., Cain, D. L., Levy, G. S., Eshleman, V. R., Fjeldbo, G., and Drake, F. D., Preliminary Results of the Mariner IV Occultation Measurement of the Atmosphere of Mars, Proc. Caltech-JPL Lunar and Planetary Conf., Pasadena, Cal., 257-266, Sept. 13-18, 1965.

Kliore, A. J., Cain, D. L., Levy, G. S., Fjeldbo, G., and Rasool, S. I., Structure of the Atmosphere of Venus Derived from Mariner 5 S-Band Occultation Measurements, Space Research, IX, Moon and Planets, North Holland Publishing Company, Amsterdam, 1969.

Kliore, A. J., Fjeldbo, G., and Seidel, B., First Results of the Mariner IV Radio Occultation Measurement of the Lower Atmosphere of Mars, Radio Sci., 5, 373-379, Feb. 1970.a.

Kliore, A. J., Fjeldbo, G., and Seidel, B. L., Summary of Mariner 6 and 7 Radio Occultation Results on the Atmosphere of Mars, Paper No. M-25, 13th COSPAR Meeting, Leningrad, U.S.S.R., 1970.b.

Kliore, A. J., Fjeldbo, G., Seidel, B. L., and Rasool, S. I., Mariner 6 and 7: Radio Occultation Measurements of the Atmosphere of Mars, Science, 166, 1393-1397, 1969.

Kliore, A. J., Levy, G. S., Cain, D. L., Fjeldbo, G., and Rasool, S. I., Atmosphere and Ionosphere of Venus From the Mariner V S-Band Radio Occultation Measurement, Science, 158, 1683-1688, Dec. 1967.

Kliore, A. J., Tito, D. A., Cain, D. L., and Levy, G. S., A Radio Occultation Experiment to Probe the Atmosphere of Venus, AIAA 5th Aerospace Sciences Meeting, Jan. 23-26, 1967. J. Spacecraft Rockets, 4, 1339-1346, 1967.

Kliore, A. J., and Tito, D. A., Radio Occultation Investigations of the Atmosphere of Mars, J. Spacecraft 4, 578-582, May 1967.

Landini, M., Russo, D., and Tagliaferri, G. L., Atmospheric Density in the 120-190 km Region Derived From the X-Ray Extinction Measured by the US NRL Satellite 1964-01-D, Nature, **206**, 173-174, 1965.

Landini, M., Russo, D., and Tagliaferri, G. L., Atmospheric Density Measured by the Attenuation of the Solar X-Rays Monitored on the NRL 1965-16-D Satellite, Icarus, **6**, 236-241, 1967.

Langworthy, B. M., HRT-11 Model Matching Program, Raytheon Tech. Memo MSD/AERO-3674, May 4, 1967.

Link, F., Eclipse Phenomena, Advances in Astronomy and Astrophysics 2, Zdenek Kopal, ed., Academic Press, 87-198, 1963.

Link, M. F., Theorie Photometrique des Eclipses de Lune, Bull. Astron., **8**, 77-108, 1932.

Link, M. F., Occultations et Eclipses par les Planetes, Bull. Astron., **10**, 227-249, 1934.

Levy, G. S., Ostoshi, T. Y., and Seidel, B. L., Ground Instrumentation for Mariner IV Occultation Experiment, IEEE Trans. Instrumentation and Measurements, **IM-16**, 100-114, 1967.

Lombardini, P. P., Doviak, R. J., and Coldhirsh, J., Centimetric and Millimetric Wave Spectroscopy of Planetary Atmospheres from Orbital Pairs, Raytheon IRP Rep. FR-66-242, Rev. A, March 1, 1967.

Lusignan, B., A Preliminary Study of Atmospheric Density Measurements by Means of Satellites, Final Report on Contract NAS 9-7020, Stanford Electronics Lab., March 1968.

Lusignan, B., Density Measurement with Radio Wave Occultation Techniques, Space Research IX, 603-609, North-Holland Publishing Co. Amsterdam, 1969.

Lusignan, B., Inverting Radio Occultation Data Using Empirical Orthogonal Function, JPL Tech. Rep. 32-1475, 115-117, July 1, 1970.

Lusigna B., Modrell, G., Morrison, A., Pomalaza, J., and Ungar, S. G., Sensing the Earth's Atmosphere with Occultation Satellites, Proc. IEEE, **57**, 458-467, 1969.

Mariner Stanford Group, Venus: Ionosphere and Atmosphere as Measured by Dual-Frequency Radio Occultation of Mariner V, Science, **153**, 1678-1683, December 29, 1967.

Martin, F. L., and Wright, F. E., Radar Ray Refraction Associated with Horizontal Variations in the Refractivity, J. Geophys. Res., **68**, 1861-1869, 1963.

Meeus, J., Occultations of Bright Stars by Planets, J. Brit. Astron. Assoc., **70**, 174-183, 1960.

Menzel, D. H., and de Vancouleurs, G., Results From the Occultation of Regulus by Venus, Astron. J., **65**, 351, Aug. 1960.

Menzel, D. H., Optical Refraction in a Planetary Atmosphere with Application to the Apparent Diameter of a Planet and to Occultation, GCA Tech. Rep. 61-33-A, Contract AF33(616)-7413, Geophysics Corp. America, July 1961.

Menzel, D. H., and de Vancouleurs, G., Final Report on the Occultation of Regulus by Venus, ARDC Contract AF19(604)-7461, July 7, 1959.

Mikhailov, A. A., Teoriya Zatmeniy (Theory of Eclipses), Moscow, 1954.

Miller, J. D. and Goldman, H. J., Feasibility of Multi-Satellite Occultation (Refraction) Measurements for Meteorology, IIT Research Institute, Tech. Memo 28, Final Report on Task 1 of Contract NASr-65(25), Aug. 1968.

Pannekoek, A., Über die Erscheinungen, welche bei einer Sternbedeckung durch einen planeten auftreten, Astron. Nachrichten, 164, 5-10, 1903.

Pettit, E., and Richardson, R. S., Motion Pictures of the Occultation of Sigma Arietis by Jupiter on November 20, 1952, Publ. Astron. Soc. Pac., 65, 91-92, April 1953.

Phinney, R. A., and Anderson, D. L., On the Radio Occultation Method for Studying Planetary Atmospheres, J. Geophys. Res., 73, 1819-1927, 1968.

Pirraglia, J., and Gross, S. H., Latitudinal and Longitudinal Variation of a Planetary Atmosphere and the Occultation Experiment, Planet. Space Sci., 18, 1769-1784, 1970.

Pomalaza, J. C., Remote Sensing of the Atmosphere by Occultation Satellites, Stanford Univ. Rep. SEL-69-045, Sept. 1969.

Rasool, S. I., and Stewart, R. W., Results and Interpretation of the S-Band Occultation Experiments on Mars and Venus, J. Atmos. Sci., 28, 869-878, Sept. 1971.

Raytheon Company, Electromagnetic Probing of the Martian Ionosphere and Atmosphere from an Orbiting Pair, Final Report on Contract NAS-8-21222, R67-4537, Dec. 19, 1967.

Raytheon Company, Mariner Mars '71 Missions: Orbiter-to-Orbiter Radio Occultation Experiment, Technical Proposal R68-4285-1, June 21, 1968.

Raytheon Company, Electromagnetic Probing of the Mars and Venus Atmospheres and Ionospheres from an Orbiting Pair, Final Report on Contract NASw-1772, R69-4347-1, 2, Sept. 19, 1969.

Shawham, S. D., Ionization Rate and Electron Density Profile for the Martian Ionosphere Based on Mariner IV Observations, 47th Annual AGU Meeting, Washington, D.C., April 19-22, 1966.

Sherr, I., Bravoco, R. R., Grossi, M., and Langevin, P. E., Profile Inversion Processing of Radio Occultation Data for the Determination of Planetary Atmospheres, Conf. on Scientific Appli-

Shklovskii, I. S., Physics of Solar Corona, Pergamon Press, Oxford, 1965.

Sodek, B. A., Jr., Jovian Occultation Experiment, J. Spacecraft Rockets, 5, 461-463, 1968.

Stanford University, SPINMAP, Stanford Proposal for an International Network for Meteorological Analyses and Prediction, Final Report, June 1966.

Stanford Electronics Lab. Proposal to NASA for a Microwave Occultation Experiment on the Nimbus E Meteorological Satellite, RL-5-68, Stanford Univ., Feb. 29, 1968.

Tatarskiy, V. I., Determining Atmospheric Density from Satellite Phase and Refraction Angle Measurements, Izv. Bull. Acad. Sci. USSR, Atmos. Oceanic Phys., 4, 699-709, 1968 a.

Tatarskiy, V. I., The Accuracy Achievable in Determining Atmospheric Density from Satellite Phase and Refraction Angle Measurements, Izv. Bull. Acad. Sci. USSR Atmos. Oceanic Phys., 4, 811-818, 1968 b.

Vitkevich, V. V., New Method of Investigating the Solar Corona, Dokl. Akad. Nauk., USSR, 77, 585-588, 1951.

Wallio, H. A., and Williams, J. R., A Feasibility Study of Determining Physical Properties of the Martian Atmosphere by use of Solar Occultation as Seen from an Artificial Satellite, NASA TN D-5443, 1969.

Wiechert, E., and Geiger, L., Bestimmung des Weges der Erdbebenwellen im Erdinnern, Phys. Z., 11, 294, 1910.

Weisberg, H. L., The Study of Planetary Atmospheres by Stellar Occultation, RM-3279-JPL (Contract N-33561 NAS 7-100), Rand Corp., Oct. 1962.

Weisberg, H. L., On the Possibility of Studying Planetary Atmospheres by Observing Stellar Occultations from a Flyby Space Probe or Planetary Orbiter, Icarus, 2, 226-227, Oct. 1963.

Werbowetzki, A., Refraction Data with Multiple Satellites, Bull. Amer. Meteorol. Soc., 47, 199, March 1966.

4. IONOSPHERIC SOUNDING

As in Chapters 2 and 3, we are concerned here with active radar sounding, both monostatic and bistatic, but at wavelengths such that the probing signals are negligibly affected by the atmosphere, but are reflected by the ionosphere. The subject is a half-century old, and quite sophisticated procedures have been developed to invert the radar data to yield profiles of electron density. The experiment has been performed only on the earth's ionosphere, although it is applicable to any planetary ionosphere. Curiously, quite different techniques and algorithms are in use for "bottomside" sounding and for "topside" sounding. They are described in the following six papers. A companion paper on VLF sounding of the D region may be found in Chapter 8. The papers all deal with pulse, or amplitude, sounding of the ionosphere; i.e., only the time delay characteristics of the radar echoes are used. More sophisticated techniques using, for example, phase, direction of arrival, or polarization have appeared in the literature but have not found extensive application. Certain of the papers in Chapter 6 on electromagnetic scattering in fact relate to this latter subject.

J. E. Jackson organized and chaired the session devoted to ionospheric sounding.

THE $P'(f)$ TO $N(h)$ INVERSION PROBLEM IN IONOSPHERIC SOUNDINGS

John E. Jackson

NASA Goddard Space Flight Center

ABSTRACT

N73-11602

A general review is given of the inversion techniques used to derive the ionospheric electron density N as a function of altitude h from group path P' versus frequency f measurements obtained by vertical incidence ionospheric sounders. The paper discusses the medium under investigation, the experimental techniques used to obtain the $P'(f)$ data, the theoretical considerations leading to the integral equation relating $P'(f)$ to $N(h)$, and the assumptions made in the inversion process. The lamination inversion technique is then presented, with special attention given to mathematical difficulties arising from discontinuities in the $P'(f)$ function, infinities in the integrand, and in some cases unknown integration limits. Methods outlined for minimizing the uncertainties due to discontinuities include the use of redundant information — that is the two distinct $P'(f)$ functions available for a given $N(h)$ profile — and the use of models based on statistical data. Mathematical procedures are discussed that increase significantly the efficiency and accuracy of the required numerical integrations. The accuracy of the inversion technique is deduced by comparing the resulting $N(h)$ profile with $N(h)$ data obtained by simultaneous but independent observations.

INTRODUCTION

Ionospheric data obtained by the vertical incidence pulse sounding technique have been used extensively to derive profiles of electron density N versus altitude h . The calculation of $N(h)$ profiles from the $P'(f)$ sounder data requires the inversion of an integral equation of the form:

$$p'(f) = \int_{p_0}^{p_r(f)} n' \left[N(p), f, B(p), \Phi(p) \right] dp \quad (1)$$

where

p' = apparent range

f = frequency

p = propagation path

p_0 = location of sounder

p_r = location of reflection point for frequency f

n' = group refractive index

N = electron density

B = amplitude of terrestrial magnetic field vector \mathbf{B}

Φ = angle between \mathbf{B} and direction of propagation

The fundamental formula for n' is:

$$n' = n + f \frac{\partial n}{\partial f} \quad (2)$$

where n , the real part of the refractive index of the medium, is given by

$$n^2 = 1 - \frac{2X(1-X)}{2(1-X) - Y_T^2 \pm \sqrt{Y_T^4 + 4(1-X)^2 Y_L^2}} \quad (3)$$

where

$$X = \frac{(fN)^2}{f^2} \quad (4)$$

$$fN = \text{plasma frequency (MHz)} = 8.98 \times 10^{-3} \sqrt{N} (\text{electrons/cm}^3) \quad (5)$$

$$Y = \frac{fH}{f} \quad (6)$$

$$fH = \text{gyrofrequency (MHz)} = 2.8B (\text{gauss}) \quad (7)$$

$$Y_T = Y \sin \Phi$$

$$Y_L = Y \cos \Phi$$

\pm = positive sign in front of square root is for the ordinary wave; negative sign is for the extraordinary wave.

The ordinary wave reflects when $X = 1$ (except when $\Phi = 0$) or equivalently [as seen from Eq. (4)] when:

$$f = fN \quad (8)$$

The extraordinary wave reflects when $X = 1 - Y$ (if $Y < 1$) and also when $X = 1 + Y$. In $P'(f)$ terminology, however, only the echo corresponding to $X = 1 - Y$ is called an extraordinary echo, and the echo corresponding to $X = 1 + Y$ is called the Z echo. Using Eqs. (4) and (6) to express the reflection conditions in terms of fN and fH yields the frequency f_x of the extraordinary echo:

$$f_x = \frac{1}{2} \left(fH + \sqrt{4fN^2 + fH^2} \right) \quad (9)$$

and the frequency f_z of the Z echo:

$$f_z = f_x - fH \quad (10)$$

Equation (1) was written in terms of a generalized propagation path p to indicate that the sounder data do not necessarily correspond to "vertical" incidence. The propagation path is, however, essentially a vertical path, if the electron density distribution is a function of altitude only (spherically stratified ionosphere). Most $P'(f)$ to $N(h)$ inversion techniques assume that stratification is essentially spherical within the ionospheric region from which the sounding echoes are received and that the propagation path is vertical. Except when otherwise indicated, these assumptions are also made here — that is, in Eq. (1) p is replaced by h and Φ is replaced by $(90^\circ - \theta)$ where θ is magnetic dip angle.

The complexity of the n' function [Jackson, 1969a] makes it impossible to evaluate analytically the integral in Eq. (1), except in very special cases ($\Phi = 0$ and $\Phi = 90^\circ$). Numerical integration techniques have to be used [even for a very simple $N(h)$ model], and special care must be taken near the reflection points where n' becomes infinite. It should also be noted that $N(h)$ is the only unknown, since B and Φ are specified by the sounder location and can be calculated to a high order of accuracy from a terrestrial field model [Cain and Cain, 1968].

The difficulties encountered in the inversion of Eq. (1) are due primarily to gaps and discontinuities in the $P'(f)$ data. The redundancy of the data [availability of more than one $P'(f)$ function] helps overcome these problems. But in many cases it is also necessary to have some knowledge of the gross features of the $N(h)$ distribution. A complete discussion of the inversion process must therefore

include some basic information concerning the medium under investigation and the limitations of the $P'(f)$ measuring technique. The characteristics of the ionosphere and of the sounder technique are such that two sounders (one on the ground and the other in an earth-orbiting satellite) are required to permit a complete determination of the $N(h)$ distribution at a given location. Here we discuss the $P'(f)$ to $N(h)$ inversion problems associated with both types of soundings.

THE MEDIUM UNDER INVESTIGATION

Since the ionosphere is produced primarily by ionizing radiation from the sun (offset by subsequent recombination and/or transport of the electron-ion gas thus created), the $N(h)$ distribution depends on the level of solar activity and undergoes diurnal, seasonal, and latitudinal variations. The terrestrial magnetic field influences not only the n' function, but also the $N(h)$ distribution. Both effects are due to the field control (through the $\mathbf{V} \times \mathbf{B}$ mechanism) on the motion (transport or radiowave-induced oscillations) of the free electrons. Since the magnetic axis is not aligned with the rotation axis of the earth, the magnetic control produces longitudinal variations in the $N(h)$ distribution (in addition to variations due to local time differences).

Although the variability of the medium can introduce some difficulties (or limitations) in the $N(h)$ calculations, certain features of the daytime and nighttime $N(h)$ distributions are sufficiently regular — at least at

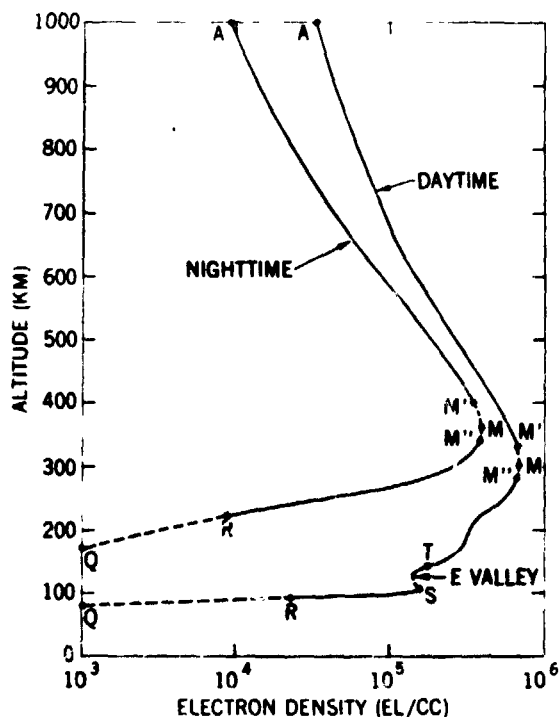


Figure 1. — Vertical electron density distributions based upon simultaneous topside and ground-based observations at Wallops Island, Virginia (latitude 37.84° N, longitude 75.47° W, $fH = 1.45$ MHz, $\theta = 69^\circ$). The daytime curve is for June 18, 1968 at 1200 EST and the nighttime curve is for June 4, 1968 at 0245 EST. The two $N(h)$ profiles corresponded to the same magnetic index ($K_p = 3$) and 10.7 cm solar flux ($S_{10.7} = 144$). The altitude difference between S and T was made about 15 km greater than observed in order to show the E valley more clearly.

midlatitudes — to provide a basis for the conventional division of the ionosphere into D, E, F1, and F2 regions. The boundaries between these regions have never been precisely defined, but it is usually understood that the D region is below 90 km, the E region between 90 and 140 km, the F1 region between 140 and 180 km, and the F2 region above 180 km. The typical features of the midlatitude $N(h)$ distribution are shown in figure 1, a daytime and a nighttime profile measured over Wallops Island, Virginia. An important characteristic of the distributions is the fact that the electron density becomes significant at an altitude of about 80 km in the daytime and about 150 km at night (points Q on fig. 1) and increases almost monotonically up to the height of maximum density ($h_{\max} F2$, shown on fig. 1 as point M), which occurs typically at altitudes between 200 and 400 km. The major exceptions to this rule occurs in daytime in the region between 110 and 140 km (E valley). Above $h_{\max} F2$ (topside ionosphere) the electron density decreases monotonically and in very nearly exponential fashion.

EXPERIMENTAL DETERMINATION OF THE $P'(f)$ FUNCTION

The ionospheric sounder, or ionosonde, operates on principles similar to those of radar, and it provides echoes from the ionosphere over a wide range of operating frequencies. Ionospheric soundings are typically conducted at frequencies between 0.1 and 20 MHz, using 100 μ sec pulses and a 20-to-60 pulse/sec repetition rate. A complete frequency sweep takes typically 15 to 30 sec.

In the widely used swept-frequency system the received echoes are customarily displayed in the "ionogram" format, in which the echo round-trip time Δt is displayed (in the vertical axis direction) as a function of the sounder frequency f . The quantity $(\Delta t)c/2$ (where c = velocity of light in vacuo) is the apparent range P' ; it is greater than the distance to the echoing region because the sounding signal is retarded by the ionosphere. The apparent range P' at a frequency f_r is therefore:

$$P'(f_r) = \frac{c}{2} (\Delta t_r) \quad (11)$$

where

$$\Delta t_r = 2 \int_{p_0}^{p_r} \frac{dp}{V_G} \quad (12)$$

$V_G = V_G(f, N, B, \Phi)$ = the velocity of the sounding signal

p_r = a reflection point for the frequency f_r and the integration is along the ray path.

Combining Eqs. (11) and (12) gives

$$P'(f_r) = \int_{p_0}^{p_r} \frac{c}{V_G} dp \quad (13)$$

The quantity c/V_G is the group refractive index n' , and Eq. (13) becomes

$$P' = \int_{p_0}^{p_r} n' dp$$

as given in Eq. (1). The vertical axis of the ionogram is calibrated in terms of P' and the basic measurement provided by the ionogram is P' as a function of frequency.

CHARACTERISTICS OF TYPICAL $P'(f)$ CURVES

The three reflection conditions [Eqs. (8) through (10)] give rise to the ordinary, extraordinary, and Z traces on ionograms. Conventional soundings from the ground rarely yield usable Z traces, but they normally provide both

ordinary and extraordinary traces. Each $P'(f)$ trace can be used (with appropriate inversion techniques) to derive $N(h)$ profiles. The ordinary trace is generally the most useful one on ground-based soundings and the extraordinary trace is the most useful one on topside soundings [Jackson, 1969a, p. 963]. In both cases (topside and ground-based soundings), the presence of at least one trace in addition to the one used for primary data analysis provides redundant information that can be used to check assumptions made in the inversion of Eq. (1) and to minimize uncertainties due to gaps in the primary data. Further redundancy of information is also available on topside ionograms, as the result of resonance phenomena that occur at fN , fH , and at the upper hybrid frequency fT , given by

$$fT = \sqrt{fN^2 + fH^2} \quad (15)$$

These resonances permit the calculation of N and B at the satellite, which is particularly helpful when the propagation traces are not clearly defined at the satellite. Since N and B (consequently, fN and fH) both decrease monotonically with altitude in the topside ionosphere, the $P'(f)$ function is known in principle for both ordinary and extraordinary traces for the complete topside profile below the satellite (portion AM of profiles in figure 1). On ground-based soundings, however, the $P'(f)$ function is not defined for the portion QR of the profiles on figure 1 (where R corresponds to the minimum density for which echoes are obtained by the sounder), and for the portion ST of the daytime profile.

The ordinary ray $P'(f)$ functions corresponding to $N(h)$ profiles of figure 1 are shown on figures 2 and 3 for topside and for ground-based soundings. To permit a direct comparison between apparent heights P' and the actual reflection heights, the $N(h)$ profiles have been redrawn with the density expressed in terms of plasma frequency [using Eq. (5)].

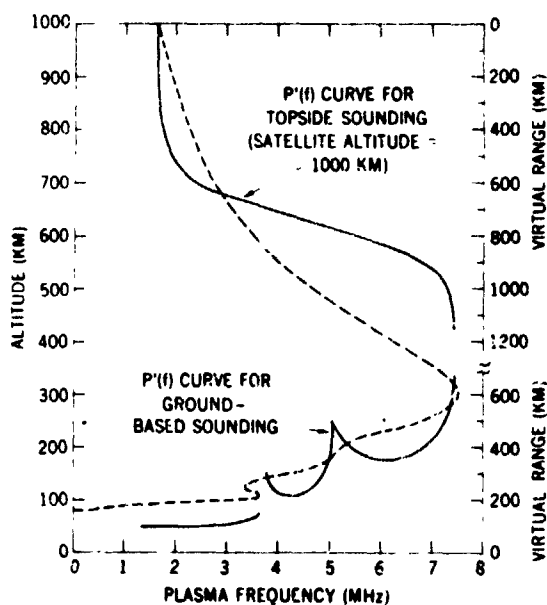


Figure 2.— Vertical distribution of plasma frequency (dashed curve) corresponding to the daytime $N(h)$ profile of figure 1, and corresponding ordinary ray $P'(f)$ functions (solid curves) for topside and ground-based soundings. The virtual range scale is one-half the altitude scale to keep the ionograms from overlapping.

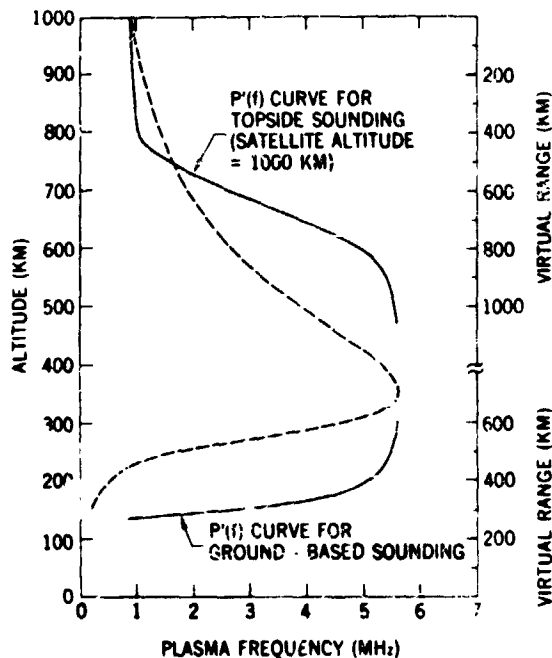


Figure 3.— Vertical distribution of plasma frequency (dashed curve) corresponding to the nighttime $N(h)$ profile of figure 1, and corresponding ordinary ray $P'(f)$ functions (solid curves) for topside and ground-based soundings. The virtual range scale is one-half the altitude scale to keep the ionograms from overlapping.

It is seen that topside and ground-based soundings terminate at $h_{\max}F2$, which follows from the monotonic variation of fN and fx with electron density. Actually, the soundings do not quite reach $h_{\max}F2$: the topside soundings stop typically at a point M' (15 to 30 km above point M of figure 1) and ground-based soundings stop typically at a point M'' (10 to 20 km below M). The region $M'M''$ can be derived from simultaneous topside and ground-based soundings [Jackson, 1969b] or by extrapolation of either type of soundings. This extrapolation usually assumes that the $N(h)$ distribution can be represented by a Chapman function in the vicinity of $h_{\max}F2$: i.e., $N(h)$ is assumed to be given by:

$$N = N_{\max} \exp \frac{1}{2} (1 - z - e^{-z}) \quad (16)$$

where

$$z = (h - h_{\max})/H$$

H = scale height

There is one additional source of information that can be used to check the total profile obtained by simultaneous topside and ground-based soundings. Topside ionograms frequently display echoes reflected from the ground (ground trace or ground echoes). These echoes, which are for frequencies greater than the maximum frequency reflected from the ionosphere, exhibit a delay that is determined by the total profile and by the frequency of the sounder signal. It is therefore possible to check the accuracy of the total profile by comparing the observed ground echoes with the ground echoes calculated from the total $N(h)$ distribution [Jackson, 1969b].

MATHEMATICAL TECHNIQUES USED IN THE INVERSION PROCESS

Here we discuss the technique used for the inversion of Eq. (1) for the case of vertical propagation — that is,

$$h'(f_r) = \int_{h_0}^{h_r} n' \left[N(h), f_r, B(h), \theta(h) \right] dh \quad (17)$$

Equation (17) is written

$$h'(f_r) = \sum_{i=0}^{k-1} \int_{h_i}^{h_{i+1}} n' dh \quad (18)$$

where

$$h_k = h_r$$

$$n' = n' \left[N(h), f_r, B(h), \theta(h) \right]$$

For each interval $\Delta h = h_{i+1} - h_i$ (or lamination), the $N(h)$ function is approximated by a simple analytic function. The number of laminations required can be minimized by choosing a function which approximates the $N(h)$ distribution over large Δh increments. A linear approximation is adequate, however, if small height increments are chosen, and this simpler representation will be used to illustrate the procedure. The linear substitution:

$$h = h_i + a_{i+1}(N - N_i) \quad dh = a_{i+1} dN$$

gives for a typical lamination:

$$\int_{h_i}^{h_{i+1}} n' dh = a_{i+1} \int_{N_i}^{N_{i+1}} n' dN \quad (19)$$

The integral corresponding to a given lamination is thus completely defined except for a constant a_{i+1} , which represents the value of dh/dN in the lamination. For all laminations, except the last one (h_{k-1}, h_k), the function n' is finite* and the integral on the right-hand side of Eq. (19) can be evaluated with great accuracy by numerical integration techniques, using, for example, a gaussian quadrature formula with only three coefficients. The last lamination terminates at the reflection point where n' is infinite. The integration in this case is based on asymptotic expressions for n' near the reflection conditions, which are of the form:

$$n' = \frac{a}{\sqrt{1 - \beta N}}$$

where a and β are constants appropriate for the propagation mode used. By letting $t^2 = 1 - \beta N$, the last integral in Eq. (19) becomes:

$$\int_{N_i}^{N_{i+1}} n' dN = -\frac{2}{\beta} \int_{t_i}^{t_{i+1}} n' t dt$$

At the reflection point, the integrand $n't$ is finite since $n't = n' \sqrt{1 - \beta N} = a$.

Most of the techniques currently used for the inversion of Eq. (1) are based on the above lamination concept. The laminations are implicitly specified by the choice of sampling frequencies, as illustrated by the following example based upon the ordinary $h'(f)$ function for a topside sounding. Let f_1, f_2, \dots, f_n be the sampling frequencies and h'_1, h'_2, \dots, h'_n be the corresponding h' values. Let N_i be the density at which reflection occurs for the frequency f_i , and let N_1 be the density at the satellite, corresponding to $h'_1 = 0$ (or equivalently $h_1 = 0$).

*There is one exception to this statement, and this occurs in the case of Z' echoes on topside ionograms where n' becomes infinite at the start of the first lamination when $X = (1 - Y^2)/(1 - Y_L^2)$.

Using the linear approximation of Eq. (19), the virtual range values can be written:

$$\begin{aligned}
 h'_2 &= a_2 \int_{N_1}^{N_2} n'(f_2) dN \\
 h'_3 &= a_2 \int_{N_1}^{N_2} n'(f_3) dN + a_3 \int_{N_2}^{N_3} n'(f_3) dN \\
 &\dots \dots \dots \\
 h'_n &= a_2 \int_{N_1}^{N_2} n'(f_n) dN + \dots + a_n \int_{N_{n-1}}^{N_n} n'(f_n) dN
 \end{aligned}
 \tag{20}$$

Each integral in Eq. (20) can be evaluated using techniques previously discussed; thus, the unknown a_i terms can be readily computed.

For the extraordinary ray the procedure is slightly more complicated, because knowledge of both f_i and the corresponding B_i at the height h_i is required to compute N_i [see Eq. (9)]. The procedure in this case is to initiate the calculation of each new lamination with an estimated value of B and to refine this estimate by an iteration process [Jackson, 1969a, pp. 973-975; Lockwood, 1970].

The laminations commonly used for $P'(f)$ to $N(h)$ inversion are of the form:

$$h = h_i + a_{i+1}(\Psi - \Psi_i) + b_{i+1}(\Psi - \Psi_i)^2 \tag{21}$$

where Ψ is n , fN , or $\ln N$. When a parabolic lamination is used ($b_{i+1} \neq 0$), the additional lamination parameters are obtained by assuming that $dh/d\Psi$ is continuous at the lamination boundaries, namely:

$$a_{i+1} + 2b_{i+1}(\Psi_{i+1} - \Psi_i) = a_{i+2} \tag{22}$$

EXPERIMENTAL CONSTRAINTS PLACED ON THE $P'(f)$ FUNCTION

Some difficulties arise in the inversion of Eq. (1) because the experimental data do not yield the $P'(f)$ function for the complete altitude range of interest. These difficulties are encountered primarily with ground-based soundings due to regions QR and ST (fig. 1) for which the $P'(f)$ functions are not available. In some cases on topside ionograms, the extraordinary $P'(f)$ function is not defined with sufficient accuracy at the starting point A . This usually arises when the density at the satellite is less than 200 electrons/cc.

Unknown Start on Topside Ionograms

If the density N_1 at the satellite altitude h_1 is unknown, it can be determined by the following method. The equations of the first two laminations are assumed to be linear in $\ln N$; that is,

$$h = h_1 + a_2 \ln \frac{N}{N_1} \quad \text{for } h_2 \leq h \leq h_1$$

and

$$h = h_2 + a_3 \ln \frac{N}{N_2} \quad \text{for } h_3 \leq h \leq h_2$$

It is then assumed that the correct value of N_1 will yield $a_2 = a_3$. Various values of N_1 are tried (by an iteration process) until a_2 and a_3 agree within 0.1%. This slope-matching technique has been tested extensively with ionograms where N_1 could be accurately measured by conventional scaling techniques. It was found that for ionograms where N_1 is well defined, conventional and slope-matching techniques have comparable accuracy. Results obtained by means of slope matching have cast considerable doubt on the determination of N_1 using the method introduced by Hagg [1967], which is based on topside sounder plasma resonance observations under conditions of low N_1 . This method yields values of N_1 that are typically one-third of the slope-matching technique values. Furthermore, the Hagg "beat" values (so-called because they are determined from the observed beat frequency between the fH and fT resonances) yield rather questionable $N(h)$ profiles (fig. 4). Lockwood (see p. 4-16) and Colin [private communication] have also used the slope-matching technique (or similar procedures) to compute N_1 and have reached similar conclusions concerning the Hagg beat.

Unknown Starts on Ground-Based Soundings

The $P'(f)$ function for a ground-based sounding can be written:

$$\int_0^{h_k} n' dh = h_Q + \int_{h_Q}^{h_R} n' dh + \int_{h_R}^{h_k} n' dh$$

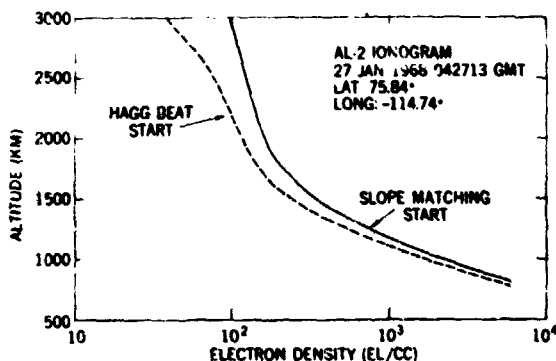


Figure 4 - Comparisons between two $N(h)$ profiles derived from the same ionogram using the Hagg-beat start (dashed-curve) and the slope matching start (solid curve).

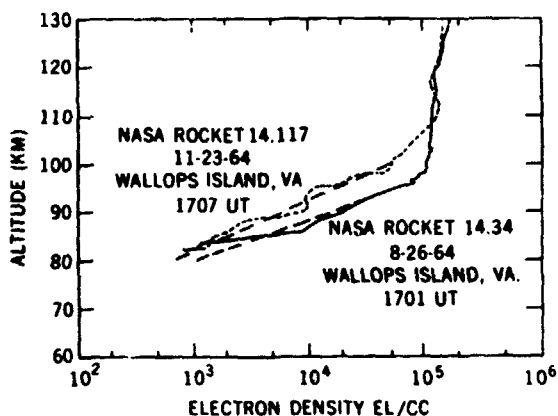


Figure 5.- Rocket results (Ref. 11) showing D and E regions. The lower part of the profile (for which the $P'(f)$ function is not normally available) can be approximated by the indicated straight segments.

where Q and R refer to figure 1, and h_k refers to a reflection point above point R . The altitude h_Q is the altitude below which n' can be considered equal to unity for the frequencies of interest. The group index n' cannot be considered equal to unity in the region QR , even though this region yields no echoes on conventional ionograms. The procedure used are different on daytime and on nighttime ionograms and these two cases will be discussed separately.

Daytime ionograms. One solution [Jackson, 1956, pp. 120-122], which is similar to the slope-matching technique, is to assume that the region QR is a simple analytic extension of the profile above point R . Another approach, based on rocket results (fig. 5), is to assume that the density is 1000 electrons/cc at 80 km, and that it increases exponentially up to point R . This gives fairly good results with midlatitude, midday ionograms.

Nighttime ionograms. Analytic extensions of the profile above point R could also be assumed for the region QR , but more reliable results can usually be obtained by making use of the ordinary and extraordinary traces, which are of comparable quality on nighttime ionograms. The procedure is to assume an arbitrary density value at Q and to compute from the ordinary trace the corresponding $N(h)$ profile. The extraordinary $P'(f)$ function corresponding to this profile is then computed and compared to the observed extraordinary $P'(f)$ function at m values of sounding frequencies. The standard deviation σ between the computed and the observed extraordinary $P'(f)$ values is

$$\sigma = \sqrt{\sum (d_i^2/m)} \quad (23)$$

where d_i represents the vertical height differences at $f = f_i$. The process is repeated for other assumed densities at Q , and the density at Q that yields the minimum value of σ is taken as the starting density. The quantity σ for typical nighttime midlatitude ionograms is shown in figure 6 as a function of the assumed plasma frequency at Q . A value of $h_Q = 150$ km was used for this calculation. Also shown is the altitude at which the electron density reaches a value of 10^5 electrons/cc. Extensive tests of this technique have given a high degree of consistency in the 150-km starting value obtained at one sounder location (Wallops Island), namely, $0.2 \text{ MHz} \leq (f/N) \leq 0.4 \text{ MHz}$. Thus an empirical starting value of $f/N = 0.3 \text{ MHz}$ could be used in cases when the two traces cannot be scaled with sufficient accuracy to permit the above determination of N_Q .

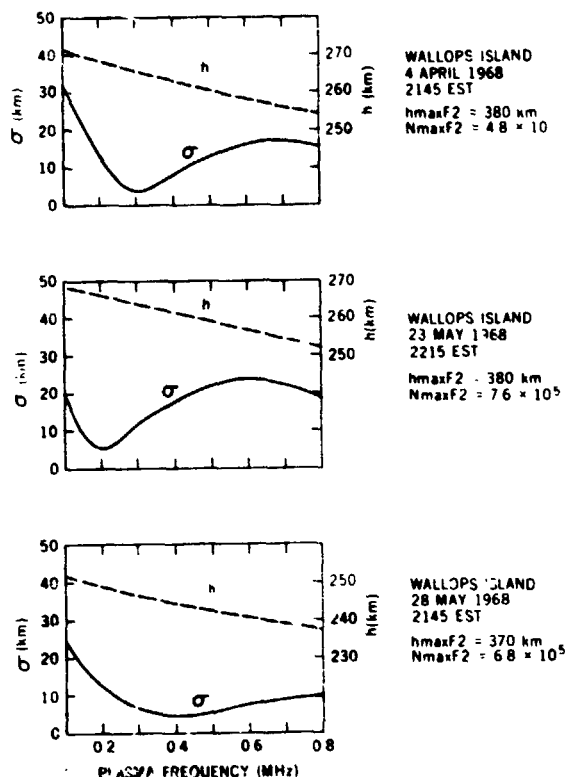


Figure 6.- The function σ versus plasma frequency at 150 km for typical midlatitude nighttime ionograms. The effect of the starting point upon the resulting $N(h)$ profile is indicated by the dashed line which shows the altitude at which the density is 10^5 el/cc.

The Valley Problem

The valley ambiguity (unknown region ST) can in principle be substantially minimized by using a technique similar to the one described for the unknown start on nighttime ground-based soundings. In this case, a model of the E valley is arbitrarily selected and the quantity σ [Eq. (23)] is computed for various valley depths. The desired valley is the one that yields the minimum value of σ . The effect of the assumed valley on the resulting $N(h)$ profile is shown in figure 7. The shape of the valley (a triangular wedge on this semilogarithmic plot) is one of many [Davis and Saha, 1962; Maliseva, 1969] assumed by other authors. Although physically unrealistic, it is convenient for

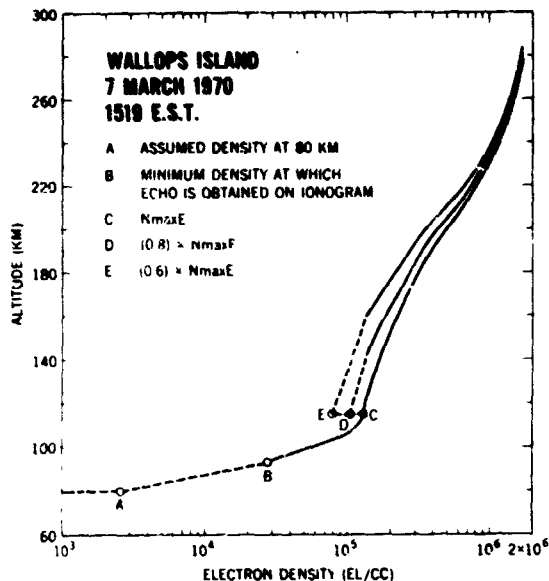


Figure 7.— Effect of assumed valley upon calculated $N(h)$ profiles

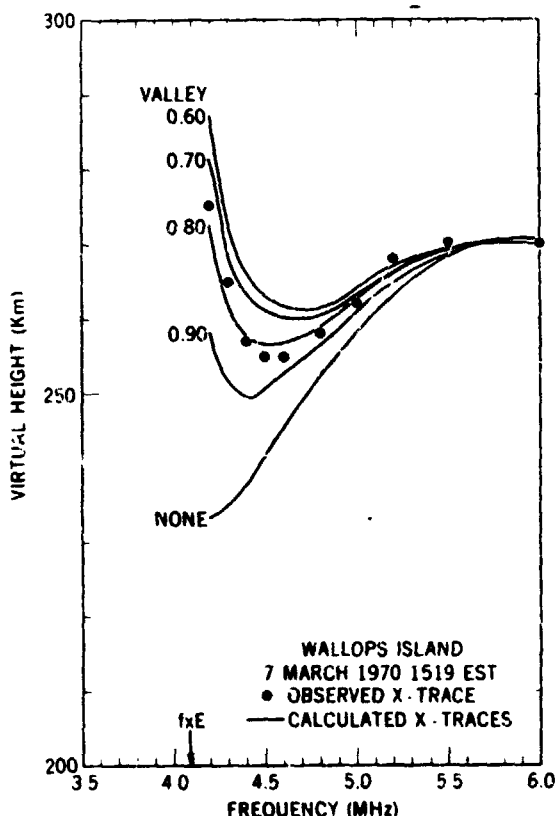


Figure 8.— Calculated X traces for $N(h)$ profiles of figure 7 and for profiles corresponding to two additional valleys (not shown in fig. 7). The observed X trace values are shown by the solid points.

sounder. These effects can be taken into consideration (see Lockwood's discussion, p. 4-26) by assuming that

computer programming. Furthermore, the shape of the valley is too variable to justify the use of a more aesthetic representation. The objective is to improve the accuracy of the profile above the valley (rather than to provide meaningful data within it). The profiles shown in figure 7 would normally be derived from the ordinary trace. These profiles and two additional ones for a 90 percent and a 70 percent valley, respectively, have been used to compute the corresponding family of extraordinary traces above E_{\max} shown in figure 8. The solid points representing the observed X trace are seen to agree best with a valley of 80 percent.

The valley determination illustrated by figures 7 and 8 was possible because the X trace was well defined for frequencies immediately above the extraordinary ray E region critical frequency $f_x E$. Often, this requirement is not met, and inspection of figure 8 indicates that the comparison would have been much less meaningful had the ionogram not shown an X trace for frequencies under, say, 4.5 MHz (particularly in view of the ± 5 -km scaling uncertainty).

The valley calculations (when feasible) yield fairly consistent results, and these results can be used to estimate the valley depth in cases when the extraordinary trace cannot be used for this purpose. For midlatitude midday conditions, both valley calculations, as illustrated above, and rocket data [Seddon and Jackson, 1958; Bauer and Jackson, 1963] show that the minimum E -valley density is typically 80 to 90 percent of the density at E_{\max} (point S of fig. 1).

PROBLEMS ARISING FROM NONVERTICAL PROPAGATION

The routine procedure for the inversion of Eq. (1) is to assume that the $P'(f)$ function represents virtual heights measured at vertical incidence. For $B = 0$, this assumption would be equivalent to stating that the sounder is receiving echoes from a spherically stratified region of the atmosphere. This region for a ground-based sounder would be a vertical cylinder typically having a diameter of 10 km. For a topside sounder, which can move horizontally up to 200 km while a complete ionogram is obtained, the spherical stratification would have to hold over a vertical region whose projection on the earth's surface would be roughly a 20- by 200-km rectangle. There is usually a change in altitude (and also some minor changes in B and θ) associated with the horizontal motion of the topside

the $N(h)$ profile is a function of altitude only and that the successive soundings begin at a new reference altitude (and corresponding density).

The spherical stratification concept usually applies to the constant electron density contours, and for $B = 0$ this assumption would yield constant phase index contours. The presence of the terrestrial magnetic field introduces modifications in the constant phase index contours, and causes the ray paths to be deviated from the vertical mostly near the reflection points. Theoretical investigations of this problem [Colin, private communication] indicate that no serious errors are introduced in the virtual height calculations by assuming that the phase index contours are the same as the electron density contours. A more important effect of the magnetic field is the occasional presence of field-aligned irregularities, which can act as wave guides and cause echoes (in the extraordinary mode) to be field aligned. This effect can be recognized by an experienced observer, and the analysis (for $\Phi = 0$) yields a field-aligned $N(h)$ profile. An example of this type of analysis is shown in figure 9

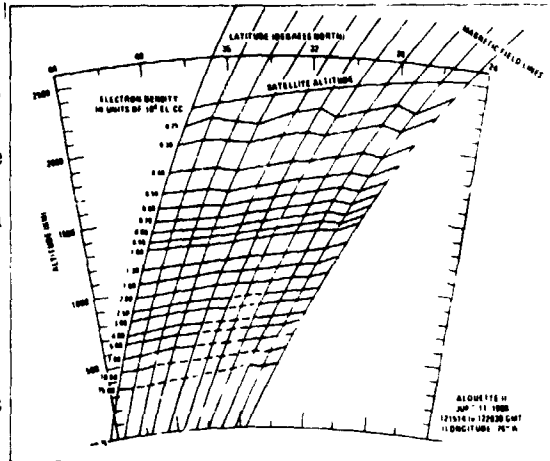


Figure 9.— Electron density contours derived from topside field-aligned traces.

Departures from spherical stratification can sometimes be inferred from a time sequence of ionograms. In the case of ground-based soundings it can be detected (when severe tilts are present) from the fact that multiple echoes (two or more round trips) do not give virtual heights in harmonic relationship to the first (one round trip) echo virtual height. Departures from spherical stratification could also be inferred from a closely spaced network of ground-based sounders. Analysis techniques that can take into consideration these departures from spherical stratification are currently under investigation (see McCulley's discussion, p. 4-36). Routine inversion techniques, however, assume that the $N(h)$ distribution corresponding to a given ionogram is a function of altitude only, and that this $N(h)$ distribution does not change significantly during the 10 to 30 sec required to obtain an ionogram.

ACCURACY OF THE INVERSION TECHNIQUES

The accuracy of the mathematical techniques used in the inversion of Eq. (1) has been tested extensively [Jackson, 1969a, pp. 967-970]. These tests have shown that the integration techniques do not introduce significant errors. For example, on topside ionograms the error in altitude for a given density is typically less than 1 km. Some errors are introduced in the inversion of Eq. (1) for ground-based soundings due to the discontinuities in the $P'(f)$ function at the starting point and at the E valley. These discontinuities introduce errors in altitude (up to 10 km) at the starting point and just above the E valley. The errors above the E valley decrease monotonically with altitude, and near $h_{\max}F2$ the calculated $N(h)$ profiles are affected by the uncertainties at the lower altitudes by at most 1 or 2 km.

Thus, it would be expected that ionograms obtained simultaneously by a ground-based ionosonde and by a topside sounder directly overhead would yield $N(h)$ profiles that should agree near $h_{\max}F2$. Comparisons of this type have yielded disagreements [Jackson, 1969b] that are several times greater than the errors attributable to scaling accuracy or the mathematical techniques used for the inversion of Eq. (1). The results suggest that the error is roughly proportional to the length of the propagation path. Since the maximum propagation path (within the ionosphere) is usually several times greater for a topside sounding than for a ground-based sounding, one would attribute this discrepancy primarily to the calculated topside $N(h)$ profile. On this basis, the topside profile appears too low by about 3 to 5% of the distance d between the topside sounder and the reflection point. One test for the accuracy of the resulting total (or composite) profile is to calculate the corresponding total content $\int_0^{h_s} N dh$ (where h_s is the satellite altitude) and compare the results with an independent measurement of the total content. Comparisons of this type using data from GEOS-2 to compute $\int N dh$ have shown consistently that the total content for the calculated

composite profile was too low. By raising the topside profile 3 to 4% of the distance d , the two measurements of $\int N dh$ can be brought into agreement. The small but puzzling discrepancy between topside and ground-based sounding suggests that some cumulative errors arise in the observations, perhaps due to irregularities in the ionosphere that cause the propagation paths to deviate from the vertical direction. The discrepancy does not seem to be due to large-scale departures from spherical stratification, as could be inferred from the analysis of ionogram sequences.

Another possibility is that cumulative errors arise in the analysis due to one of the many assumptions made in the magnetoionic theory such as cold plasma treatment of the ionosphere, *WKB* approximation, idealized reflection conditions, and group velocity representation of the signal velocity. Although these approximations have been accepted for several decades, a complete evaluation of their effects has not been made in the topside ionosphere.

SUMMARY

The $P'(f)$ to $N(h)$ inversion in ionospheric soundings is based on the following assumptions:

1. Spherically stratified medium (except in special techniques under development that make allowance for gradients).
2. Medium that does not change during the sounding.
3. Vertical propagation (except in more elaborate techniques used, for example, to deal with geomagnetic field-aligned propagation).
4. Simple ray theory of wave propagation in the ionosphere.

The most accurate inversion technique is based on:

1. Lamination concept and step-by-step solution.
2. Lamination model optimized for representative $N(h)$ distributions.
3. Numerical integration with change of variable to keep the integrand finite.
4. Iteration (with topside extraordinary $P'(f)$ function) to find the upper limit of integration.

Special problems arise, particularly with the $P'(f)$ function obtained by ground-based soundings, when the $P'(f)$ function is not defined over the complete altitude range of interest. This leads to (1) the starting point problem, and (2) the valley problem (for daytime ground-based soundings).

The solutions to these special problems are based on:

1. Redundancy of the data (ordinary and extraordinary $P'(f)$ functions).
2. Estimated parameter values derived from statistical data accumulated from previous $P'(f)$ to $N(h)$ inversions, or from other techniques, such as in situ rocket measurements.
3. In some cases, extrapolations or slope-matching techniques.

The accuracy of the $N(h)$ profiles derived by inversion of the $P'(f)$ function is typically better than 5% of the length of the propagation path in the ionosphere. The errors in the $N(h)$ profiles are apparently due to assumptions made in the analysis, since they are several times greater than the error attributable to the mathematical techniques used in the inversion process. Although the human element was not stressed in this paper, it is an important factor in the ultimate accuracy achieved with a given inversion technique: It enters into the initial recognition and accurate scaling of the $P'(f)$ function, the evaluation of the validity of the assumptions made for a given experimental condition, and the final acceptance (or rejection) of the results given by the inversion process.

THE CALCULATION OF ELECTRON DENSITY PROFILES FROM TOPSIDE IONOGRAMS - METHOD AND APPLICATIONS

G. E. K. Lockwood

Communications Research Center

Shirley Bay, Ottawa, Canada

N73-11603

ABSTRACT

A method for converting topside sounder ionograms into topside electron density profiles is discussed. The lamination method used is modified to take into account the variation of electron density and magnetic field within each lamination. Also included is a change of variable to produce a finite integrand of the integral involved, an iteration scheme that permits convergence on an initially unknown density, a second iteration scheme to overcome the problem of an uncertainty in the electron density at the satellite, and a modification to compensate for the changing satellite altitude over the duration of the sounding. Two applications of the technique are discussed: field-aligned traces for computing field-aligned profiles and computer-aided systems for scaling ionograms.

INTRODUCTION

Many schemes for converting topside ionograms into electron density profiles have been devised during the years since the launch of the first topside sounder in the Alouette 1 satellite. The most accurate and most widely used method has been described in detail by Jackson [1969]. The technique starts with the simple concept of the lamination method of inversion, which is equivalent to the inversion of a triangular matrix. Then it incorporates several complicating modifications, all of which are needed to obtain the desired accuracy of the result. These modifications include a second-order logarithmic model for the variation of height with electron density within each lamination, an allowance for the change in magnetic field within each lamination, a change of variable to keep the function being integrated finite, an iteration permitting convergence on an initially unknown limit of integration, an iteration scheme to overcome the problem of an inaccurate starting electron density, and a procedure to account for the changing satellite altitude.

Here we summarize the main points of topside $N(h)$ reduction with emphasis on how our knowledge of the ionosphere has influenced the particular solution to the problem, and discuss two special applications of the technique: the use of field-aligned ionograms for computing $N(h)$ profiles and computer-aided systems for scaling ionograms.

COMPUTATION OF TOPSIDE $N(h)$ PROFILES

Topside ionograms [Hagg et al., 1969; Jackson, 1969] are the range versus frequency records obtained from a topside sounder [Franklin and Maclean, 1969]. Topside ionograms have many varied forms and features; the ionogram shown in figure 1 was selected to illustrate clearly the Z-, O- and X-wave traces. Also identified is the Z' trace [Calvert, 1966] and various plasma resonances [Hagg et al., 1969]. Jackson has discussed the convention (followed here) of expressing echo time delay as "apparent range" in kilometers.

The Z-, O-, or X-wave trace may be used independently to compute an electron density $[N(h)]$ profile, although the Z-wave trace, because of its cutoff, usually provides only a partial profile. Figure 2 from Jackson [private communication] shows $N(h)$ profiles computed from the Z-, O-, and X-wave traces in figure 1. The profiles were computed by the inversion of the integral

$$p'(f) = \int_{h_0}^{h_r(f)} n'[N(h), f, B(h), \theta(h)] dh \quad (1)$$

where

$p'(f)$ = apparent range at the wave frequency f

h_0 = height of the sounder (satellite altitude)

h_r = height at which reflection occurs for wave frequency f

n' = group refractive index

f = wave frequency

B = magnitude of the earth's magnetic field

θ = angle between the earth's magnetic field-line and the wave normal

N = the electron density

The problem of calculating $N(h)$ from the curve $p'(f)$ scaled from a topside ionogram is discussed below.

Topside Lamination Method

In the basic lamination method, the integral is broken up into many parts, each being determined explicitly by the points (p', f) selected from the ionogram. Initially, both h_0 and N_0 are known; then:

$$p'(f_1) = \int_{h_0}^{h_1} n' dh = \int_{N_0}^{N_1} n' \frac{dh}{dN} dN \quad (2)$$

$$p'(f_2) = \int_{h_0}^{h_1} n' dh + \int_{h_1}^{h_2} n' dh \quad (3)$$

$$p'(f_j) = \int_{h_0}^{h_1} n' dh + \dots + \int_{h_{j-1}}^{h_j} n' dh \quad (4)$$

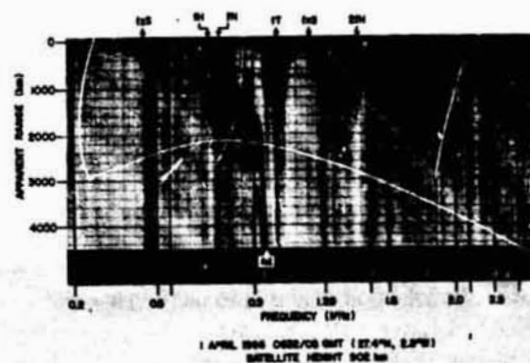


Figure 1.— Example of an Alouette II ionogram with clear Z-, O-, and X-wave traces.

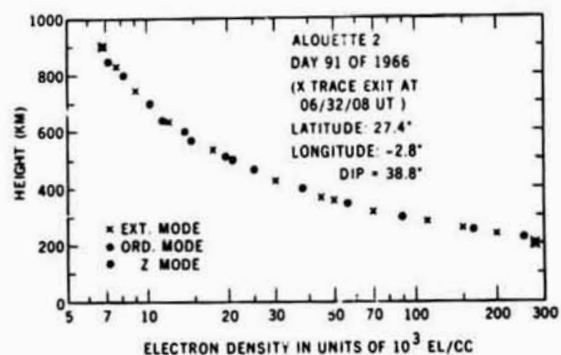


Figure 2.— Electron densities obtained from the ionogram in figure 1 with an independent $N(h)$ calculation for each of the Z-, O-, and X-wave traces. The points shown in the graph were selected from the computed points.

From Eq. (2) one finds h_1 , from (3) one finds h_2 , and from (4) one can obtain h_j . But the evaluation of $\int n'(N, f, B, \theta) dh$ is not straightforward since h_j and N_j are unknown (although $N_{j-1}, h_{j-1}, B_{j-1}$ are known) and both B and N change within the integral. Thus, models that define N and B within the lamination are needed before one can numerically determine the integral of n' .

$N(h)$ Model

Since the electron density in the topside ionosphere varies approximately exponentially with height, a suitable choice for the definition of height with density within the lamination between points $j-1$ and j is:

$$h = h_{j-1} + a_j \log \frac{N}{N_{j-1}} + \left[b_j \log^2 \frac{N}{N_{j-1}} \right] \quad (5)$$

If the term in brackets is neglected, one has a "linear in $\log N$ " model, which, though simple, has a discontinuity at the edge of each lamination. The term in brackets is necessary if the derivative is to be made continuous; the result is a "parabolic in $\log N$ " model. The term

$$a_{j+1} = a_j + 2b_j \log \frac{N_j}{N_{j-1}} \quad (6)$$

is the "scale height" at the point j and is derived from the condition that the derivative be continuous at each boundary. This is a meaningful physical parameter since the scale height is a measure of the electron and ion temperature $T_e + T_i$ and mean ionic mass \bar{m} . For an ionosphere in diffusive equilibrium, the scale height is given by:

$$H_j = -a_{j+1} = \frac{T_e + T_i}{\frac{\bar{m}g}{k} + \frac{d}{dh}(T_e + T_i)} \quad (7)$$

where k = Boltzmann's constant. This model for electron density has the advantages that dN/dh is continuous over the complete profile; the scale height, a pertinent physical parameter is implicit in the model, and the model is monotonic in N , as is the electron density in the topside ionosphere.

Magnetic Field Model

The magnetic field at any known height and location can be calculated to high accuracy from a modern spherical harmonic model such as the International geomagnetic reference field [Cain and Cain, 1968]. Thus, the field B_{j-1} at the top of the j th lamination is known, and its value at other heights within the lamination can be calculated to sufficient accuracy by assuming an inverse cube variation with height. Jackson [1969] has combined these two models in such a way that the magnetic field is defined as a function of electron density.

Change of Variable

As the wave propagates downward to its reflection height the group velocity V_g of the wave approaches zero, and since $n' = c/V_g$, the integrand n' approaches infinity. The following change of variable [Shinn and Wale, 1952;

[Jackson, 1956] is introduced to produce an integrand that is finite at reflection and permits a more accurate numerical evaluation of the integral:

$$t^2 = 1 - X \text{ for the } O \text{ wave} \quad (8)$$

$$t^2 = 1 - \frac{X}{1 - Y_R} \text{ for the } X \text{ wave} \quad (9)$$

where

$$X = fN^2/f^2$$

$$Y_R = fH_R/f$$

fH_R = gyrofrequency at the height of reflection

Numerical Integration

Generally, the accuracy of a numerical integration may be determined by two methods: (1) An analytic evaluation of the integral (if possible) provides the correct value of the integral; and (2) the number of points used in the evaluation of the numerical integration can be increased until no change occurs in the result. Such comparisons [Jackson, 1969] have shown that provided the above change of variable is performed, a three-point gaussian numerical integration produces an error of less than 0.1% for the X-wave, and an error of less than 0.1% for the O-wave, provided θ is greater than 20° . For the O wave, the percentage error increases as $\theta \rightarrow 0$, and a 16-point gaussian numerical integration is required to produce an error less than 0.1% when $\theta = 5^\circ$.

Iteration

For the lamination in which reflection occurs, the lower limit in height and its gyrofrequency (or B) are unknown. For the X wave the density is unknown also, since it is a function of both the wave frequency and gyrofrequency. One can make some initial guess for the gyrofrequency at the bottom of the lamination, perform the integration to determine the height h_j , and calculate the actual gyrofrequency at this height. Then one repeats the integration and continues the iteration until the new value of h_j is the same as the old value. When the X-wave trace is scaled, such an iteration is absolutely necessary since the guess for fH determines the estimate for the electron density at the bottom of the lamination. Jackson [1969] has shown that a useful first guess for the gyrofrequency at reflection fH_R is that corresponding to some minimum permissible increase in density over that at the top of the lamination. For example,

$$fx_j [fx_j - fH_R (\text{guess})] = 1.001 fx_{j-1} (fx_{j-1} - fH_{j-1}) \quad (10)$$

implies a 0.1% increase in density at the j th point over that at the $j-1$ th point.

An example of such an iteration is shown in figure 3. The initial guess of fH_R , calculated from the above expression, leads to convergence at the desired height at point E. Lockwood [1970] has shown that the above iteration will diverge for a restricted class of ionograms even though a solution exists. An ionogram of this class is shown in figure 4. The iteration for the first lamination in this ionogram is shown in figure 5. The iteration

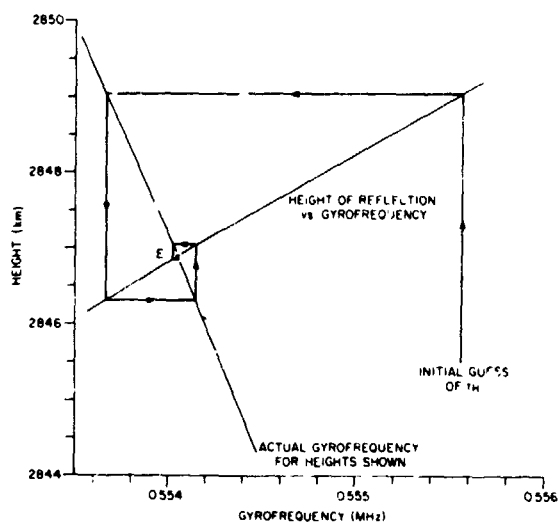


Figure 3.— An illustration of the iteration technique that converges on the desired intersection point E from an initial guess for fH_R corresponding to a 0.1 percent increase in density over the density at the top of the lamination.

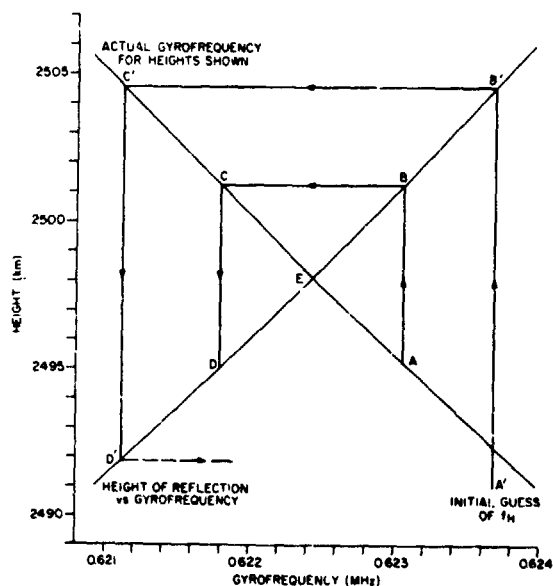


Figure 5.— An example of an iteration ($A'B'C'D'$) which diverges and a modified technique which converges to the desired point E . In the modified technique, an initial guess A leads to the points B , C , and D . A linear interpolation using points A , B , C and D result in the approximate height E . The same steps are repeated using the height E and the actual gyrofrequency at that height as a new starting point.

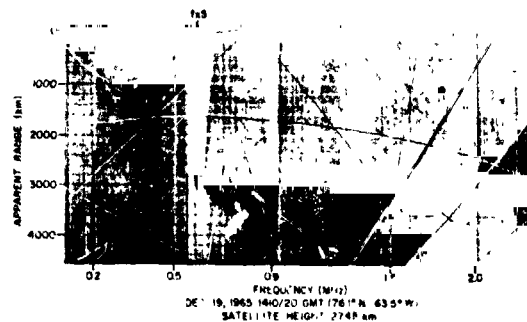


Figure 4.— An ionogram in which the density at the satellite is small. The difference frequency, $fT - fH$, is observed as a modulation on the combined fH , and fT spikes.

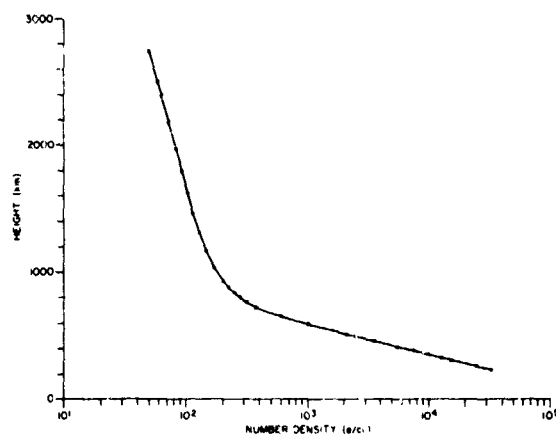


Figure 6.— The $N(h)$ profile computed from the scaled points in the ionogram of figure 4. The point at 2498 km is the result of the modified iteration shown in figure 5.

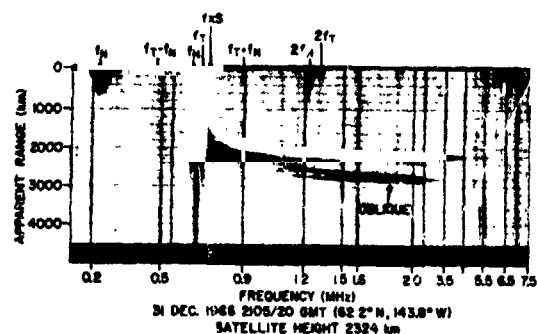


Figure 7.— An ionogram which illustrates the broad fN spike at low electron densities, and the merging of the fXs with fT and fH at low densities.

$A'B'C'D'$ diverges because the slope of the height of reflection curve (versus assumed gyrofrequency) has a larger absolute value than that of the slope of the actual gyrofrequency-height curve. A modification of the iteration, which always converges if a solution exists, is as follows:

1. Perform the standard iteration for two steps only to define the points B , C , and D from the starting point A .
2. If the difference in height between points B and D is less than some small value, the iteration is terminated and point D is taken as the correct solution.
3. Otherwise, perform a linear interpolation by using the lines AC and BD to find the approximate intersection point E . This height and the actual gyrofrequency at this height are then used as the coordinates of a new starting point in step 1. For this modified iteration, an initial guess for fH_R corresponding to an increase in density of 10 percent was used since this increase [Jackson, 1969] is typical of many scaled ionograms, and this guess permits a slightly more rapid convergence. The resultant profile is shown in figure 6.

The Starting Density Problem (For Densities Less Than 1000 Electrons/cc)

As the electron density at the satellite approaches zero, the fxS approaches fHS since $fNS^2 = fxS$ ($fxS - fHS$). The relative accuracy to which fxS can be scaled decreases as the density approaches zero since fxS merges with the gyrofrequency resonance of the ionogram (see figure 7). A starting density (or fxS) that is too small produces a "knee" in the $N(h)$ profile, in which the slope of the first lamination is smaller than the average slope of subsequent laminations. From a knowledge of the ionosphere, one expects the scale height (or slope) to be constant or to increase with height. Thus, an incorrect starting density produces an obvious error in the $N(h)$ profile.

Some possible schemes to circumvent this problem are:

1. The plasma frequency fN can be scaled to determine the starting density [Jackson, 1967]. This could provide a somewhat more accurate measurement of density but has two disadvantages. As shown in figure 7 the fN spike on the ionogram is usually broad in frequency with a corresponding uncertainty in its nominal center frequency. Also, since current topside sounders begin at 100 kHz, the minimum observable plasma frequency is 100 kHz, corresponding to an electron density of 124/cc.
2. Hagg [1967] has shown that as the electron density approaches zero, the fH and fT resonances can be simultaneously excited by the sounder transmitter pulse. The difference frequency between the two resonances is observed on the ionogram as a modulation on the single spike produced by the two resonances, an example of which is shown in figure 4. An electron number density can be calculated by assuming that each resonance occurs at its nominal frequency. Then:

$$fN^2 = 2fH \Delta f \quad (11)$$

where Δf = beat frequency. However, Lockwood and Jackson [private communication] have found that the starting density calculated by this method is too small, since it produces an $N(h)$ profile whose first scale height is smaller than the second. The starting density must be increased by a factor of 2 or 3 to produce an $N(h)$ profile in which the first scale height is at least as large as the second. The error in the calculation of fN by this method may be caused by the apparent shift of the fH and fT resonances from their nominal frequencies [Benson, 1969; Bitoun et al., 1970]; on the other hand, the calculated value may represent the actual density in the proximity of the satellite and not the average density at that altitude. Nevertheless, until the fH and fT resonances are better understood, this method cannot be used to calculate the starting density for $N(h)$ reduction.

3. Since the scale height is expected to be constant or to increase with altitude, one could try several values of fxS and select the resulting profile that appears to be correct [Jackson, 1967]. More specifically, for an isothermal ionosphere in diffusive equilibrium [Eq. (7)] the scale height at different altitudes will be the same except for the variation due to the changing gravitational field. The geopotential scale height

$$H_j' = H_j \left(\frac{R_0}{R_0 + h_j} \right)^2 \quad (12)$$

in which the gravitational change is removed, will not change with altitude under the above conditions. For a realistic $N(h)$ profile, the geopotential scale height, H_1' at the bottom of the first lamination should be equal to the geopotential scale height, H_2' at the bottom of the second lamination. Thus, one could perform the $N(h)$ calculation for the first and second laminations only, and then test to see whether $|H_1' - H_2'| < \epsilon$. If the inequality is satisfied then the $N(h)$ calculation is continued. Otherwise, if H_1' is less than H_2' , N_0 is increased by some increment and the calculation is started again, if H_1' is greater than H_2' , N_0 is decreased by some increment and the calculation started again. This technique need not be limited in use to low electron densities; it could be used for any ionogram in which the accuracy of scaling fxS or fN is poor. At CRC the iteration is performed if N_0 is less than $1000/\text{cm}^3$, with $\epsilon = 50$ km.

Changing Altitude of the Sounder

For topside sounders that are not in a circular orbit (Alouette II and ISIS-1), the change in the satellite altitude during the measurement of a single ionogram can be of importance when high accuracy is required [Eccles, private communication]. Typically, for these satellites, the maximum vertical velocity is about 1 km/sec. Thus, an error of up to 10 km is introduced by not taking into account the changing altitude of the sounder over the duration of the trace on the ionogram. Often, this error is no more significant than the uncertainties resulting from the scaling of the ionogram. But for a well-defined ionogram, the accuracy of the $N(h)$ profile is improved by the correction described below.

The change in the satellite altitude is

$$\Delta h = \Delta f \frac{dh_s/dt}{df/dt} \quad (13)$$

where $\Delta f = f_j - f_0$. The modified starting height for the profile is

$$h_0' = h_0 + \Delta h \quad (14)$$

The density at this height

$$N_0' = N_0 \exp \left(\frac{h_0' - h_0}{a_1} \right)$$

is derived from an extrapolation of the first lamination ($h = h_0 + a_1 \log N/N_0$) to h_0' . The point (h_0', N_0') instead of the point (h_0, N_0) is used as a limit of the first integral in Eq. (4).

SPECIAL APPLICATIONS

Field Aligned $N(h)$ Reduction

Several workers including *Colin, Hojo, Jackson* [private communications] and this author have shown that if one uses the X -wave field-aligned trace to compute an electron density profile along a magnetic field line, one obtains densities that agree with densities from vertical topside profiles at locations where the two profiles intersect. Thus, the capability exists for mapping the electron density in the direction of the magnetic field, as well as the vertical direction.

The modifications that are required to change the vertical $N(h)$ reduction program to a field-aligned $N(h)$ reduction program are:

1. The angle θ is changed from 90° -DIP for vertical propagation to $\theta = 0^\circ$ for field-aligned propagation.
2. For each lamination, the increment ($\Delta S = \Delta h = h_j - h_{j-1}$) is treated as a step in the direction of the earth's magnetic field instead of the vertical direction.

Muldrew [1963] has performed field-aligned ray tracing by using an electron density model containing a field-aligned discontinuity; he showed that as the energy propagated along the field, the angle θ remained less than 10° except at reflection. In the use of the field-aligned $N(h)$ reduction program, no changes result in the field-aligned profile as θ is changed from 0° to 10° . This justifies the use of the value $\theta = 0$ in computing field-aligned profiles.

A Semiautomatic Scaling System

An ionogram contains more information than the minimum required to compute an $N(h)$ profile. In principle either the O - or the X -wave trace could be used to calculate the $N(h)$ profile. The other trace could then be used to check the validity of the interpretation and scaling of the ionogram. *Lockwood* [1969] has described such a computer-assisted scaling system in which an operator, in conjunction with a computer, performs the scaling of the X -wave trace of an ionogram and the calculation of the $N(h)$ profile and corresponding O -wave trace, all in real time. A difference between the calculated and actual O -wave trace on the ionogram would indicate an error in scaling or a misinterpretation of the traces on the ionogram. The operator can rescale the ionogram until a satisfactory agreement is obtained between the actual and calculated O -wave traces. Figure 8 shows an incorrect scaling and figure 9, a correct scaling of an ionogram and its corresponding computed O -wave trace. In figure 8 the error in the scaled X -wave trace at 1.6 MHz produces a corresponding error in the O -wave trace at 1.0 MHz.

More recently, both *Lockwood and Colin* [private communication] have developed similar, more sophisticated semiautomatic systems, with more features to aid the operator in the interpretation of the ionogram. For example, at high latitudes, the vertical reflection traces are often masked by spread echoes. However, by redisplaying the ionogram on the CRT at a higher amplitude threshold level, the spread is removed from the display (but not from the original data) and the reflection traces become clearly visible, as shown in figures 10 through 12. Figure 10 shows a conventional ionogram with severe spread echoes. The same ionogram, as displayed on the computer scaling system at CRC is shown in figure 11. Figure 12 shows the digital ionogram redisplayed at a higher threshold level; in it most of the spread echoes are removed and the main traces are discernible.

A trained operator now performs the identification of the traces in the ionogram and scales the appropriate trace. The operator is better suited than a computer to perform these tasks since they involve subjective decisions in pattern recognition. On the other hand, a computer could determine the range and frequency coordinates of a trace (i.e., scale it) if the trace were identified to the computer. In the computer-aided system at CRC, the computer

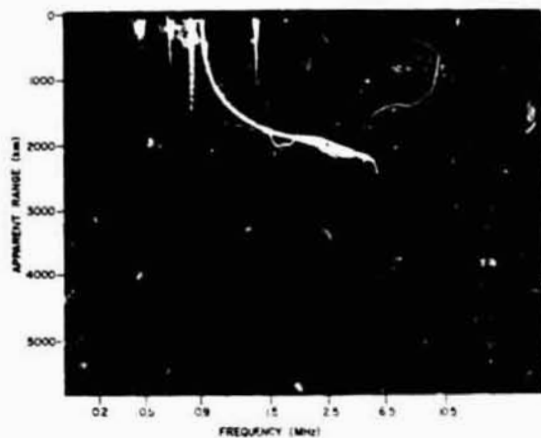


Figure 8.— A digitized ionogram showing the operator's scaled *X*-wave trace, and the calculated *O*-wave trace. The operator's error in scaling at 1.6 MHz produces an error in the calculated *O*-wave trace at 1.0 MHz.

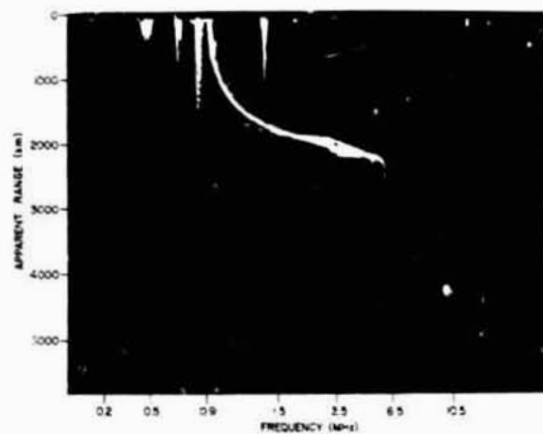


Figure 9.— The same ionogram with a revised scaled *X*-wave trace and its corresponding calculated *O*-wave trace.

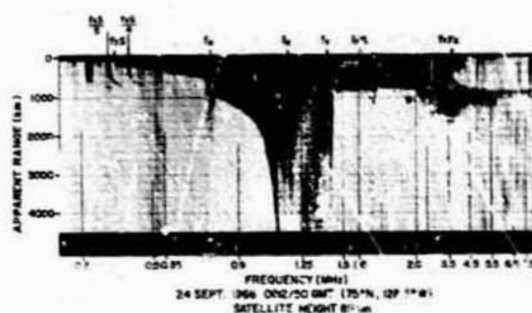


Figure 10.— A high-latitude Alouette II ionogram exhibiting severe spread echoes.

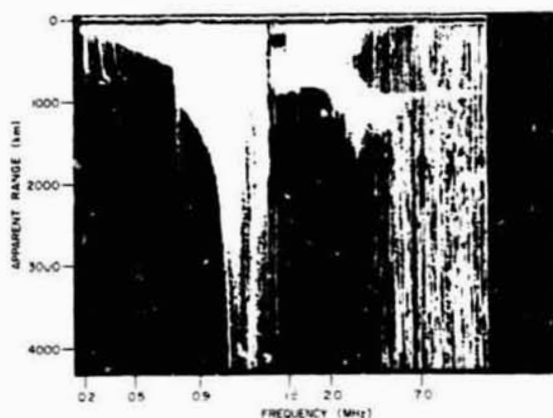


Figure 11.— The same ionogram, displayed by the computer-aided scaling system, with a display threshold level equivalent to that in the ionogram in figure 10.

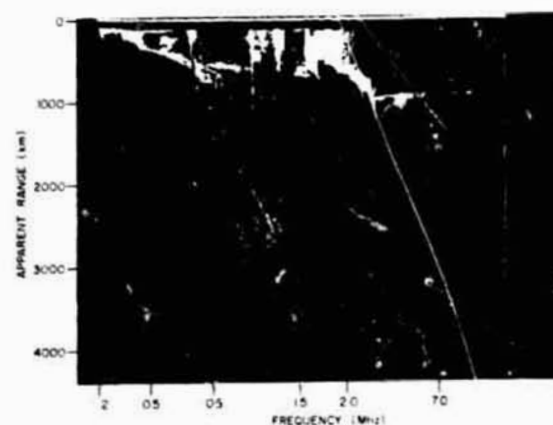


Figure 12.— The same digitized ionogram displayed at a higher threshold level to remove most of the spread echoes.

determines the center of the trace by making use of the amplitude information in the ionogram. However, computer identification of the traces has not yet been attempted because of the complexity of the problem.

Although computer identification of patterns has been applied successfully in other disciplines (the analysis of cloud chamber photographs and chromosome identification) the task is more difficult in the analysis of topside sounder ionograms for several reasons. The signal-to-noise ratio is variable and often poor. The traces are often incomplete and fragmentary. There are no distinct classes of ionograms; instead, there is a continuum of ionograms from one "type" to another. Furthermore, the identification of the vertical traces is often complicated by more prominent oblique and field-aligned traces. Thus, the computer identification of traces in topside ionogram presents a challenge that has not yet been met.

SUMMARY

This paper has illustrated how our knowledge of ionospheric physics has aided the development of a mathematical method for converting topside ionograms into electron density profiles. The scheme makes use of the fact that the electron density in the topside ionosphere varies exponentially with height with a slowly changing scale height. A spherical harmonic model is used to determine accurately the earth's magnetic field components at lamination boundaries, and an inverse cube variation in B is assumed within each lamination. Then, an iteration based on the constancy of geopotential scale height over a small height range is used to overcome an error caused by an uncertainty in the starting density.

Two special applications of this inversion scheme have been discussed: computing an electron density profile along a magnetic field line, and a semiautomatic scaling system in which redundant information in the ionogram is used to verify the interpretation and scaling of the ionogram. With regard to the second application, the computer identification of traces in topside ionograms remains a challenging problem.

DISCUSSION

Wright: Both you and Dr. Jackson referred to obtaining profiles from field-aligned traces. How do you recognize when this occurs?

Lockwood: Sometimes you see both vertical and field-aligned traces on the topside ionogram, particularly on the extraordinary trace.

Wright: Aren't there intermediate cases where there is only one echo of a kind, perhaps partly guided?

Lockwood: Yes, but you can normally recognize that. Sometimes at high latitudes you will see only the field-aligned traces.

Wright: Concerning the question of correcting geomagnetic field with altitude, you showed a rather elaborate iteration scheme that seemed to be concerned with removing an error of the order of 3 km in 3000 km of real height. In comparison with other errors do you believe this is justified?

Lockwood: Yes, because that's an error in each lamination and it is cumulative.

Jackson: In the analysis of ground-based soundings you wouldn't be concerned with this.

Croft: Are there any plans for putting a sounder in a geostationary satellite?

Lockwood: We would like to. We have proposed both a magnetosphere sounder and a sounder in a geostationary orbit.

Grossi: Concerning mirroring echoes, the topside sounder shows them to be absent above 6.5 MHz. On the contrary on the ground a recent experiment shows conjugate point echoes up to 12 MHz. How stable is the PRF of the Alouette and ISIS sounders, such that we could do some time integration from period to period?

Lockwood: I don't have a number for that, but it should be quite stable.

Grossi: Could you integrate coherently for minutes?

Lockwood: No, I don't think so for two reasons: The sounder is sweeping at the same time; and we are getting the echoes because of irregularities aligned along the field, and the satellite passes through the center of a duct and will move out in time.

Grossi: Sometimes there are large shells, however.

Unidentified speaker: Combining the topside and ground-based sounding, could you measure the valley between E and F region?

Lockwood: I don't think we would be sensitive enough to look at anything but gross features using the ground echoes.

Unidentified speaker: Have you compared topside sounding with ground-based incoherent scatter?

Lockwood: The incoherent scatter, in general, gives comparable data. These backscatter sounder do require sounder data to calibrate the F region peak density.

Jackson: Cohen, at Boulder, made 20 comparisons of Jicamarca data with simultaneous Alouette passes, and there was a systematic difference between the two, indicating that the Alouette profile was lower, the same sort of thing I mentioned before.

PRECEDING PAGE BLANK NOT FILMED

NUMERICAL METHODS FOR REDUCTION OF TOPSIDE IONOGRAMS

L. McCulley

Informatics Inc., Palo Alto, California

N73-11604

ABSTRACT

Several alternative methods for solving the group height equation are presented. Three of these are now in operation at Ames Research Center and use data contained in a single ionogram trace. From the data an electron density profile $N(h)$ is computed. If the ionogram also exhibits other traces, "reverse" ionogram traces are computed, using the $N(h)$ profile, for comparison with the redundant data. When agreement is poor, the initial data trace is reinterpreted, another $N(h)$ profile computed, and the reverse traces generated once again. This process is repeated until a desired degree of consistency is achieved.

To reduce the necessity for human intervention and eliminate decision making required in conjunction with the preceding methods, a method is proposed that accepts as input, *all* data from a single ionogram. In general, no electron density function will satisfy these data exactly, but a "best" $N(h)$ profile can be computed.

Finally, a method is described that eliminates the need to assume that the ionosphere is spherically stratified. Horizontal gradients in electron density are detected and accounted for by processing several ionograms from the same satellite pass simultaneously. This idea is derived as an extension of one of the basic methods.

INTRODUCTION

Sounder ionogram data $h'(f)$ can be converted to electron density versus altitude profiles by inverting the group height equation

$$h'(f) = \int_h^{h_1} n'(X, Y, \phi) d\bar{n} \quad (1)$$

where

h' = virtual height or depth (km)

f = sounder frequency (MHz)

h_1 = height of sounder (km)

h = height of wave reflection (km)

n' = group refractive index [Davies, 1965]

$X = (fN/f)^2 = N/(12400 f^2)$

fN = electron plasma frequency (MHz)

N = electron density (electrons/cm³)

$$Y = fH/f = 2.8B/f$$

fH = electron gyrofrequency (MHz)

B = earth's magnetic field strength (gauss) [Cain and Cain, 1968]

ϕ = angle between magnetic field vector [Cain and Cain, 1968] and direction of signal propagation (degs.)

Except for electron density N , which is assumed to be a monotonic decreasing function of altitude h , all the above variables can be considered as input to the problem.

Generally, the inversion problem is attacked by modeling the $N(h)$ profile as an analytic function depending on several free parameters. These parameters are then adjusted iteratively to satisfy both the reflection condition

$$X_r = \begin{cases} 1 - Y_r & X \text{ trace} \\ 1 & O \text{ trace} \\ 1 + Y_r & Z \text{ trace} \end{cases} \quad (2)$$

and the group height equation for each of a discrete set of data points (h', f) .

All reduction methods discussed here take advantage of the essential "triangularity" of Eq. (1) to compute only a portion of the $N(h)$ profile at each step. This is done by dividing the ionosphere into a number of spherical layers and relating altitude and electron density within each of these laminations by

$$h = h_{j-1} + a_j \ln \frac{X}{X_{j-1}} + b_j \left(\ln \frac{X}{X_{j-1}} \right)^2 \quad (3)$$

Coefficients a_j and b_j are then computed for each lamination in turn.

Although each layer model has two parameters, only one is left free. In the topmost lamination, b_2 is set to 0 so that only a_2 is free. In subsequent laminations, a_j are computed by

$$a_j = a_{j-1} + 2b_{j-1} \ln \frac{X_{j-1}}{X_{j-2}} \quad (4)$$

leaving b_j free. Equation (4) provides for continuity in the first derivative of $N(h)$ at the boundaries h_j . [Note that continuity of $N(h)$ itself is implicit in the model given by Eq. (3).]

Jackson was the first to present in detail a method for solving this problem [Jackson, 1969a]. However, the iteration scheme used is open ended (not constrained) and as a result, convergence is not always obtained. More recently, Lockwood [1970] has improved the iteration method, but convergence still depends on the accuracy of an initial estimate to the solution.

One of the primary considerations in development of the Ames direct, inverse, and field-aligned reduction methods was elimination of uncertainties in both accuracy and convergence in the calculations. As a result, convergence within a specified error tolerance can be guaranteed, if only a solution exists.

Each of these basic methods operates only with data from a single trace (X , O , or Z) of a single ionogram. The direct and inverse methods assume a vertical (radial) ray path and are distinguished from each other only by a difference in control variable. Wave propagation is assumed to be directed along the earth's magnetic field lines in the field-aligned method.

Inverse processing is extended to cope with the problem occurring when significant horizontal (latitudinal or longitudinal) gradients exist in $N(h)$ — that is, when the ionosphere cannot be accurately represented by a layer model. This same basic method is also shown to have application in computing a "best" $N(h)$ profile for all data from a single ionogram.

THE DIRECT AND FIELD-ALIGNED METHODS

The Direct Method

A set of points (h_j', f_j) is chosen to represent the ionogram trace. Reflection heights h_j corresponding to the signal frequencies f_j are taken to be the lamination boundaries and as such are a natural control variable for the iteration process. The direct reduction method computes these boundaries.

It should be observed that with this one to one correspondence between data points and lamination boundaries, data selection has a direct effect on the accuracy provided by the $N(h)$ model. Too few data points can give a poor result regardless of how smooth the ionogram trace may be.

Reflection heights h_j are found by solving for the root of the group height function

$$F(h_j) = h'(f_j) - \int_{h_j}^{h_1} n'(X, Y, \phi) dh \quad (5)$$

This is where the direct method differs from the iteration schemes used by Jackson and Lockwood.

This solution process is necessarily iterative, requiring that the function F be evaluated several times for successively more refined estimates of h_j . Since f_j is specified, an estimate of h_j determines a value for F according to the following sequence of operations:

Calculation	Source
$h_j \rightarrow h_j'$	the magnetic field parameters are known
compute ϕ_j	the magnetic field parameters are known
$Y_j \rightarrow X_j$	the reflection condition, Eq. (2)
compute a_j	Eq. (4) (computed only once and only for parabolic laminations)
$h_j, X_j \rightarrow b_j$	Eq. (3) (parabolic laminations)
$\rightarrow a_j$	Eq. (3) (linear laminations)
compute $F(h_j)$	Eq. (5)

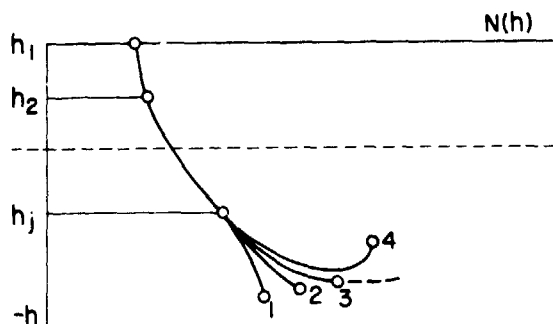


Figure 1.- Direct Method - $N(h)$ laminations for four reflection heights h_{j+1} . Curve 3 corresponds to the maximum allowable value since $N(h)$ must be single valued.

formed carefully. If h_j is chosen too large, the parabolic $N(h)$ model will be forced to take a minimum in the h direction as shown by curve 4 in figure 1. This does not physically occur in the topside ionosphere and will be reflected by an imaginary value of n' when evaluated at points within the lamination.

An upper limit for h_j is found by solving $dh/d(\ln X) = 0$ where h is given by Eq. (3). Calculation of this maximum involves solution of a system of equations since the reflection condition and magnetic field relation must also be satisfied. Curve 3 in figure 1 represents the model profile corresponding to this marginal solution. Curves 1 and 2 represent possible acceptable solutions.

Observe that in Eq. (5), only the integral over the interval (h_j, h_{j-1}) need be recomputed as h_j is varied.

There are many available processes by which the value of h_j can be refined toward the root of the function F . The one used here is a modified version of the Regula Falsi method [Ostrowski, 1960]. Upper and lower bounds on h_j are determined. Intermediate values can then be selected in such a way as to provide a sequence of sets of bounding values that are successively more restrictive. When the root can be bounded initially, convergence can be guaranteed.

Since F cannot be evaluated for an arbitrarily specified value of h_j , the initial boundary process must be per-

For a discussion of the lower limit on h_j , the ionogram traces must be considered separately.

X trace: The reflection condition $X = 1 - Y$ provides an upper limit for $Y_j = 1 - X_{j-1}$. The minimum value of h_j , then, is the value corresponding to the magnetic field strength given by $\max Y_j$.

O trace: No physical limit can be inferred. An arbitrary value of 10 km is prescribed.

Z trace: The quantity $X - (1 + Y)$ must remain negative throughout the lamination if Z wave propagation is to be allowed. Curve 3 in figure 2 corresponds to a physically impossible solution. Curve 2 represents the marginal case corresponding to the solution of $d[X - (1 + Y)]/dh = 0$. Curve 1 corresponds to a possible acceptable solution. As before, the reflection condition and the magnetic field relation are imposed.

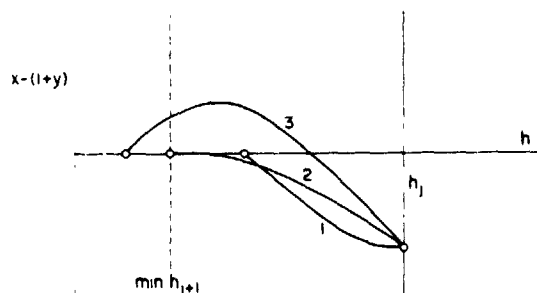


Figure 2.- Direct Method - Behavior of $X - (1 + Y)$ for three reflection heights h_{j+1} (Z-trace calculations only). Curve 2 corresponds to the minimum allowable value since $X - (1 + Y)$ must be negative.

Limits computed in this way will normally bracket the root h_j to be computed; for example, the sign of the function F will be different for evaluations at the upper and lower limits. However, if no sign change is observed then no root exists, and it becomes pointless to initiate the iteration process. Furthermore, since the lack of a root generally indicates an error in the data, the value of F obtained (h_j' is only a reference level to this value) indicates the correction to the data which is necessary so that a solution to Eq. (5) does exist.

The Field-Aligned Method

In most cases, the results of data reduction as performed assuming vertical signal propagation, are consistent with the known physical situation. Sometimes, however, a vertical assumption is obviously invalid and the most likely alternative is that the signal has propagated along "tubes" aligned with the earth's magnetic field.

A field-aligned method is designed for data reduction when such is assumed. In concept this method is identical to the direct one, and requires only that h and h' be replaced by distances p and p' , along the field line discussed above, and that the calculations be performed along that field line.

THE INVERSE METHOD

Inverse processing is so designated because instead of specifying frequencies and computing reflection heights as with the direct method, one solves the inverse problem. Frequencies are determined that correspond to reflection at a specified set of altitudes. Inverse processing is also discussed by *Madsen*, [1970a].

One advantage of this approach is that the approximation error inherent in the $N(h)$ model can be more easily controlled. More significant, however, are the sophisticated extensions of the method that are possible.

The ideas that motivated the calculation process of the direct method also prevail here. A group height function

$$G(f_j) = h'(f_j) - \int_{h_j}^{h_1} n'(X, Y, \phi) dh \quad (6)$$

is defined, and the root f_j found by iteration. The following sequence of operations indicates how the function G is evaluated for a given estimate of f_j .

<u>Calculation</u>	<u>Source</u>
compute B_j, ϕ_j	the magnetic field parameters are known (computed only once)
$f_j \rightarrow Y_j$	$Y_j = 2.8 B_j / f_j$
$Y_j \rightarrow X_j$	the reflection condition, Eq. (2)
compute a_j	Eq. (4) (computed only once and only for parabolic laminations)
$h_j, X_j \rightarrow b_j$	Eq. (3) (parabolic laminations)
$\rightarrow a_j$	Eq. (3) (linear laminations)
$f_j \rightarrow h'_j$	An initial set of data points (h', f) is "curve-fit", and h'_j corresponding to f_j is found by interpolation.
compute $G(f_j)$	Eq. (6)

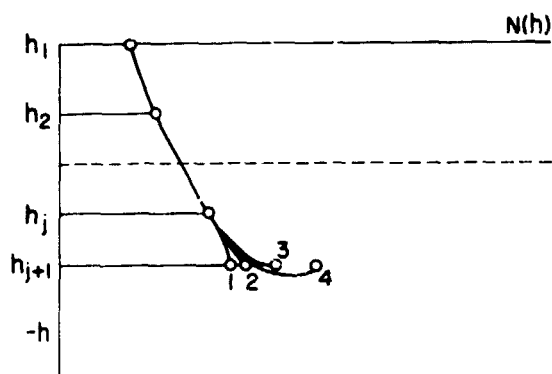


Figure 3.— Inverse Method — $N(h)$ laminations for four frequencies reflecting at h_{j+1} . Curve 3 corresponds to the maximum allowable frequency since $N(h)$ must be single valued.

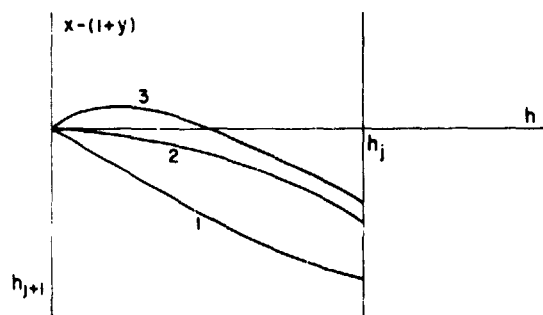


Figure 4.— Inverse Method — Behavior of $X-(1+Y)$ for three reflection frequencies (Z-trace calculations only). Curve 2 corresponds to the minimum allowable frequency since $X-(1+Y)$ must be negative at every point within the lamination.

entire altitude range for each trial value of f_j . An added complication, though not as severe, is in providing the capability to numerically determine a data value h'_j for any specified f_j .

THE OPTIMAL LAMINATION METHOD

This method attempts to profitably use, simultaneously, all the available ionogram data pertinent to the calculation of a single lamination of the $N(h)$ profile. At least one of the ionogram traces must be visible beginning at zero virtual depth. The inverse method then proceeds as usual with the exception that instead of searching for the root of the group height function G [Eq. (6)], one looks for the minimum of a given object function. The purpose of this function is to provide a measure of the consistency between the data and the computed $N(h)$ profile.

Consider that the reflection heights h_2, h_3, \dots, h_N (and thus the gyrofrequencies fH_2, fH_3, \dots, fH_N), have been specified. The inverse method then computes the corresponding frequencies f_2, f_3, \dots, f_N .

For a given estimate of N (or fN) at the current reflection height h_j , a frequency f_j is obtained for each of the ionograms traces from the reflection conditions

As before, the bounds are obtained from the $N(h)$ model and the reflection condition. Curve 4 in figure 3 corresponds to a physically impossible solution. An upper bound on f_j is found by solving $dh/d(\ln X) = 0$ in conjunction with the proper reflection condition and the magnetic field relation. Curve 3 in figure 3 represents the maximum attainable electron density and thus by the reflection condition, a maximum value for f_j . Curves 1 and 2 correspond to possible acceptable solutions. In this case, the maximum can be directly evaluated. It is not necessary to solve a system of equations.

Computation of a lower limit again depends on the trace from which the calculations are performed:

X trace: The reflection condition is solved for $\min f_j$ with $fH = fH_j$ and $fN = fN_{j-1}$

O trace: $\min f_j = f_{j-1}$

Z trace: the quantity $X-(1+Y)$ must remain negative throughout the lamination. Curve 3 in figure 4 corresponds to a physically impossible solution. Curve 2 represents the marginal case corresponding to $d[X-(1+Y)]/dh = 0$. Curve 1 corresponds to a possible acceptable solution.

Although the inverse method does offer some important advantages over the direct method, it is inherently more time consuming. This follows primarily from the necessity for computing the integral in Eq. (6) over the

$$X \text{ trace: } X = 1 - Y \rightarrow fN^2 = f(f - fH) \rightarrow f = (fH + \sqrt{fH^2 + 4fN^2})/2$$

$$O \text{ trace: } X = 1 \rightarrow fN^2 = f^2 \rightarrow f = fN$$

$$Z \text{ trace: } X = 1 + Y \rightarrow fN^2 = f(f + fH) \rightarrow f = (-fH + \sqrt{fH^2 + 4fN^2})/2$$

Using these frequencies, one then computes from Eq. (1), the corresponding virtual height

$$h'_c(f_j) = \int_{h_j}^{h_1} n'(X, Y, \phi) dh$$

for each trace. The subscript c indicates a computed value. The object function, defined for example as

$$J(fN) = \sum_{X, O, Z} [h'_c(f_j) - h'(f_j)]^2$$

is minimized with respect to fN . Note that if $h'(f_j)$ is visible for only a single trace, the minimum value of $J(fN)$ is zero and the method is equivalent to the inverse.

As an example, consider the case where the X trace is continuously visible from the satellite position but not out so far as to display ground echoes; the visible O trace begins at some nonzero virtual height and does exhibit ground echoes; the Z trace is partially visible (fig. 5). All visible data would be scaled and curvefit for later interpolation.

If the altitudes h_1, h_2, \dots, h_7 correspond to the data as shown in figure 5, then the $N(h)$ profile (fig. 6) would be determined in the first lamination by the scaled X and Z traces, in the second lamination by the X trace alone, in the third and fourth laminations by the X and O traces, and in the remaining laminations by the O

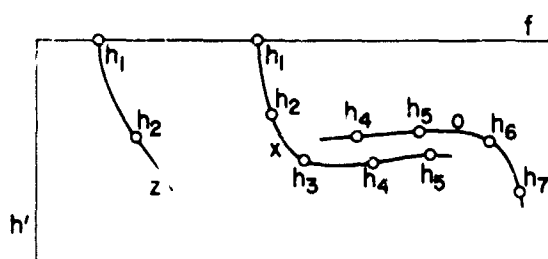


Figure 5.— Idealized sample ionogram showing visible portions of X , O , and Z traces. The points h_j indicate the altitudes at which the corresponding frequencies have reflected.

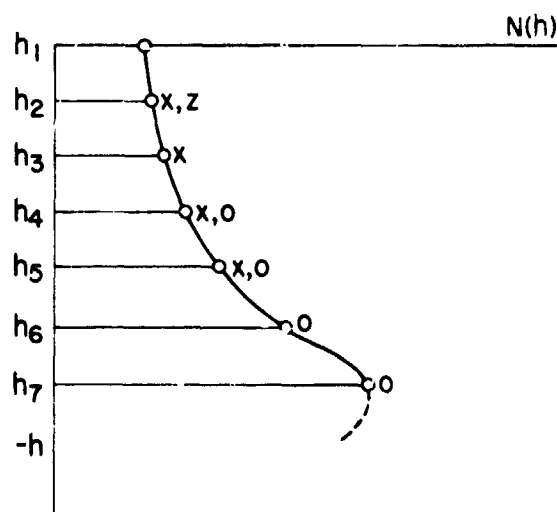


Figure 6.— Idealized $N(h)$ profile as would be computed by the optimal lamination method. Data from the indicated traces is used in the computation of the individual lamination boundaries (see fig. 5).

trace alone. Thus, all visible ionogram data impact the solution for the $N(h)$ profile in a single, computer-controlled operation.

Since the choice of the object function J is arbitrary, various bits of scaled data can be weighted according to the confidence which they inspire. Also, a different object function can be selected with only minor changes to the computer program.

THE HORIZONTAL GRADIENT METHOD

A fundamental feature of ionograms produced by orbiting swept frequency sounders is that each of the transmitted pulses travels through a different ionosphere due to satellite motion within the orbital plane. The reduction methods discussed previously do not take this into account, but assume that the ionosphere is spherically stratified. This is not generally true. The existence of regions where the ionosphere displays significant horizontal gradients in electron density are well known [Jackson, 1969b].

A conceptually simple method [see Madsen, 1970b for a more detailed description] that accounts for this horizontal variation of electron density can be employed if two or more ionograms are processed simultaneously. The basic computing block of this method is the inverse method above.

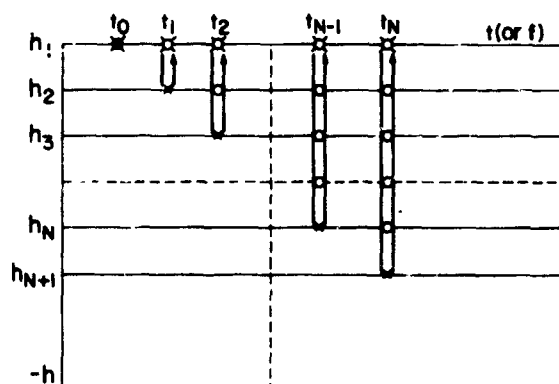


Figure 7. - Reflection heights versus time. The x locations are points where electron density is calculated by the Inverse Method. The o locations are points where interpolated values of electron density would be obtained in the Horizontal Gradient Method.

Consider a point in the orbital plane to be defined by the coordinates h = reflection height and t = time (frequency could be used in place of time). For specified reflection heights h_2, h_3, \dots, h_{N+1} , the inverse method computes $N(h)$ lamination profiles sequentially, downward from the satellite position. Each successive calculation occurs at a later time and requires as input, the values of N at all reflection heights in figure 7 directly above the one currently considered.

For example, the calculation of $N(h_3, t_2)$ requires as input the values $N(h_2, t_2)$ and $N(h_1, t_2)$. However, only the approximations $N(h_2, t_1)$ and $N(h_1, t_0)$ are known from previous calculations when ionograms are processed singly. Using these approximate values in the calculations is equivalent to assuming spherical stratification in the ionosphere.

If the horizontal gradients in $N(h)$ were known, this variation with time could be incorporated into the calculations to give more accurate results. Specifically, if values of $N'(h) = (dN/dt)_h$ for $h = h_1, h_2, \dots, h_j$ are available, then in the calculation of $N(h_{j+1})$, one could use the following more accurate values of N at t_j :

$$\begin{aligned} N(h_1, t_j) &= N(h_1, t_0) + (t_j - t_0)N'(h_1) \\ N(h_2, t_j) &= N(h_2, t_1) + (t_j - t_1)N'(h_2) \\ &\dots \dots \dots \\ N(h_j, t_j) &= N(h_j, t_{j-1}) + (t_j - t_{j-1})N'(h_j) \end{aligned} \quad (7)$$

The horizontal gradient method establishes values for the derivatives $N'(h)$ by simultaneous calculation of several contiguous ionograms. The following step-by-step procedure describes the method.

1. Compute $N(h_1, t_0)$ for each ionogram.
2. Define $N'(h_1)$ to be the slope of the "best" straight line through these points when N is plotted versus time.
3. Using Eq. (7), compute the values $N(h_1, t_1)$ for each ionogram.
4. Using the inverse method, compute $N(h_2, t_1)$ for each ionogram.
5. Repeat steps 2 through 4 for $h = h_2, h_3, \dots, h_{N+1}$

The first-order variations of $N(h)$ in the time direction are thus accommodated.

Theoretically, it is possible to detect and measure horizontal gradients within the time (distance) necessary to produce a single ionogram, through the judicious use of resonances, electromagnetic cutoffs, and electron gyrofrequency resonances. In practice, however, this approach is precluded by the inaccuracies due to resonance broadening and, in some cases, poor frequency resolution [Colin, private communication].

One suspects then that it might be possible, over a distance of several ionograms, to account for more detailed variation of $N(h)$ in the horizontal direction. This hypothesis could be tested by computing the two parameters describing the "best" parabolic description of the horizontal variation in step 2 above, and modifying Eq. (7) accordingly. A good parabolic fit to the data would justify this approach.

SUMMARY

The direct, field-aligned, and inverse methods for ionogram reduction operate reliably and as accurately as possible on data obtained from a single ionogram trace. The inverse method is extended to provide an $N(h)$ profile consistent with all data from a single ionogram. This is accomplished without operator intervention. A separate extension of inverse provides greater accuracy in the computed $N(h)$ profile. This is accomplished by processing several ionograms simultaneously so as to account for horizontal gradients in electron density.

In the description of the methods for which signal propagation is vertical, the sounder height h_1 has been assumed constant, for simplicity only, throughout the calculation. The direct and inverse methods as implemented at Ames Research Center do account for sounder altitude variations. This capability can also be incorporated into the proposed horizontal gradient method with little effort.

The field-aligned method fixes the sounder at some initial position so that the time-consuming process of computing field line parameters is not required for each individual data point. It follows that development of inverse and horizontal gradient analogs for field-aligned signal propagation is not feasible, since they would be even more complicated and time consuming.

DISCUSSION

Parker: I didn't fully understand your method for taking out the assumption of spherical layering. Do you need many ionograms to do this?

McCulley: You need at least two.

Wright: The solution to a succession of ionograms obtained along the path of the satellite isn't necessarily going to resolve the horizontal gradient unless the gradient also lies along the satellite's path.

McCulley: These satellites are mostly in polar orbits where latitude is the main change along the satellite track.

Croft: When you make soundings of the ionosphere using techniques that are sensitive to the map of the ionosphere, like *HF* ground backscatter, you see much evidence of traveling irregularities that have periodic structure. You would have to be careful that your sweep rate wasn't in synchronism with this period.

THE MATHEMATICAL-PHYSICAL PROBLEM OF $p'(f)$ TO $N(h)$

INVERSION FOR ANISOTROPIC MEDIA

A. K. Paul

National Oceanic and Atmospheric Administration

Boulder, Colorado

N73-11605

ABSTRACT

It is well known that the calculation of electron density profiles from ionograms cannot provide unique results, since the information required is incomplete. The anisotropy of the ionosphere provides a means to reduce the uncertainty, as has been recognized by *Argence*, [1954] and *Storey* [1960]. A systematic investigation about the independence of the two magnetoionic components show that the extent to which the range of ambiguity can be reduced depends mainly on the magnetic latitude. On the other hand, it is even more important that the numerical method used for the inversion is basically very accurate. Studying the properties of the integral transform we come directly to some conclusions about the minimum requirements for a numerical inversion method. Some remarks are given about error estimates and first-order corrections for less elaborate methods.

INTRODUCTION

In vertical sounding of the ionosphere we obtain the virtual height h' as a function of radio frequency f . The virtual height is related to the reflection height by an integral equation of Abel's type. If this integral equation can be solved we obtain first the reflection height as a function of the radio frequency; then, using the reflection condition relating radio frequency and electron density at the reflection level, we obtain the electron density as a function of height, the electron density profile. The solution process of the integral equation requires that the virtual height be known for all frequencies up to the frequency for which the reflection height is to be determined and that the electron density be a monotonic increasing function of height. While the second condition may be fulfilled for a relatively large percentage of ionograms, the first condition is practically never fulfilled [*Manning*, 1949]. To obtain an exact and unique solution, we need additional information. Some additional information is available as will be shown later, but usually it is not sufficient [*Paul and Smith*, 1968]. We then have to introduce a physically reasonable model assumption and try to estimate the range of uncertainty of the profile obtained this way. Some insight into the errors of this process can be obtained by studying the properties of the integral transform. At the same time we will learn about the requirements for a rapidly convergent method of numerical solution [*Paul*, 1967].

Since the ionosphere is a doubly refracting medium, the question arises if and how much additional information can be obtained by using both polarizations in the inversion process. The results of an investigation of this problem will be presented later on.

We propose some practical procedures giving first-order corrections, when no additional information is available or where computing facilities are very limited.

PROPERTIES OF THE INTEGRAL TRANSFORM

The virtual height h' is given by

$$h' = \int_0^{h_R} \mu' dz \quad (1)$$

where h_R is the height of reflection and z is the height variable. If we neglect the earth's magnetic field, the group refractive index is

$$\mu' = \frac{1}{\sqrt{1-X}} \quad (2)$$

where $X = f_N^2/f^2$, f is the radio frequency, f_N the plasma frequency and $f_N^2(z)$ is proportional to the electron density. The radio signal is reflected where $X = 1$ or $f_N(h_r) = f$. If the electron density is never decreasing with increasing height, the virtual height can be expressed by

$$h'(f) = h_0 + f^2 \int_0^1 \frac{1}{\sqrt{1-X}} z' dX \quad (3)$$

where h_0 is the lower boundary of the ionosphere and $z' = dz/df_N^2$.

For a given electron density profile Eq. (3) represents an integral transform. Its most essential feature is the singularity at the reflection height. Equation (3) can be rewritten

$$h' = h_R + f^2 \int_0^1 \left(\frac{1}{\sqrt{1-X}} - 1 \right) z' dX \quad (4)$$

We see immediately that the difference between virtual height and reflection height is determined mainly by the slope of the profile near the reflection height, since the factor by which z' is weighted in Eq. (4) is very small for small values of X and tends to become infinite when X approaches unity. Similarly, it is easy to see that h' is always finite if z' is finite. On the other hand, it is possible that h' is finite, even if z' is infinite. This can be demonstrated in the following example: Consider an electron density profile given by

$$f_N^2 = f_c^2 \left[1 - \left(\frac{z - h_m}{y_m} \right)^{2r} \right] \quad (5)$$

For $r = 1$ this represents a parabolic profile with the maximum density proportional to f_c^2 at the height $z = h_m$ and a half thickness y_m . If we solve Eq. (5) for z and differentiate, we obtain as the expression for the slope

of the profile

$$z' = \frac{y_m}{f_c^2} \frac{1}{2r} \left(1 - \frac{f_N^2}{f_c^2} \right)^{-(2r-1)/2r} \quad (6)$$

If $r > 1/2$ the slope becomes vertical for $f_N = f_c$, since z' becomes infinite. Introducing Eq. (6) into (3), the integral is convergent for $f = f_c$ if $r < 1$, and we obtain

$$h'(f_c) = h_0 + \frac{1}{1-r} y_m \quad \text{for } \frac{1}{2} < r < 1 \quad (7)$$

For $1/2 < r < 1$ we therefore have a model layer with a maximum density for which the virtual height at the critical frequency is finite. The parabolic maximum with $f = 1$ represents just the limiting case between finite and infinite virtual height as is indicated by the logarithmic singularity in the expression for the virtual height of a parabolic layer. The profile and the virtual height for $r = 2/3$ are shown in figure 1. A few more examples may illustrate how different the effect of a vertical slope on the virtual height can be. A model given by

$$z(f_N) = \begin{cases} h_0 & 0 \leq f_N \leq f_m \\ h_0 + y_m \left[1 - \frac{2}{\pi} \arcsin \left(\frac{f_m}{f_N} \right) \right] & f_N > f_m \end{cases} \quad (8)$$

is continuous at the base of the layer where $f_N = f_m$, but it has an infinite slope. The virtual height is given by

$$h' = \begin{cases} h_0 & 0 \leq f < f_m \\ h_0 + y_m & f > f_m \end{cases} \quad (9)$$

There the vertical slope of the profile results in a finite discontinuity of h' . This is shown in figure 2. The virtual heights shown here are usually interpreted as reflections from two flat thin layers at different heights with relatively low electron density in between, which is probably correct, but cannot be proven by using h' observations alone, as demonstrated in this example.

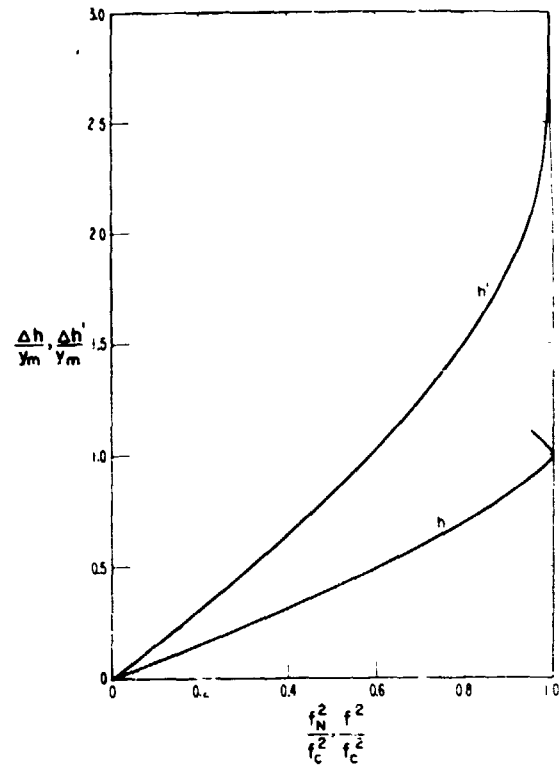


Figure 1.— Electron-density profile (labeled h) and virtual-height curve (h') for the model given by 2.5 and 2.7.

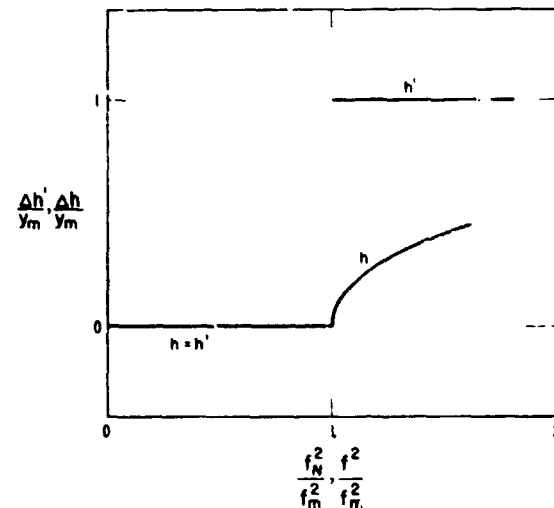


Figure 2.— Profile (h) and virtual-height curve (h'), on normalized scales, for the model given by 2.8 and 2.9.

For models given by

$$z(f_N^2) = \begin{cases} h_0 & 0 \leq f_N^2 \leq f_m^2 \\ h_0 + y_m \left(\frac{f_N^2}{f_m^2} - 1 \right)^{k/2} & f_N^2 \geq f_m^2 \end{cases} \quad (10)$$

again we have a continuous profile and an infinite z' at $f_N = f_m$, if $k < 2$. The virtual height, however, may have an infinite discontinuity or a finite discontinuity, or it may be continuous, depending on whether $0 < k < 1$, $k = 1$, or $1 < k < 2$. Profiles and virtual heights for the three cases are shown in figure 3.

The examples so far have shown that a continuous profile may have a continuous h' curve even if z' is infinite, or a singularity of z' may result in a finite or infinite discontinuity of μ' . On the other hand, if a finite or infinite discontinuity is observed in an ionogram it is impossible to find out whether the profile is continuous or not.

It is also interesting to study the effect of a discontinuity of z' . In figure 4 we have a model where the electron density is linearly increasing with height, but the slope is discontinuous at $f_N^2 = f_1^2$. In the upper portion the

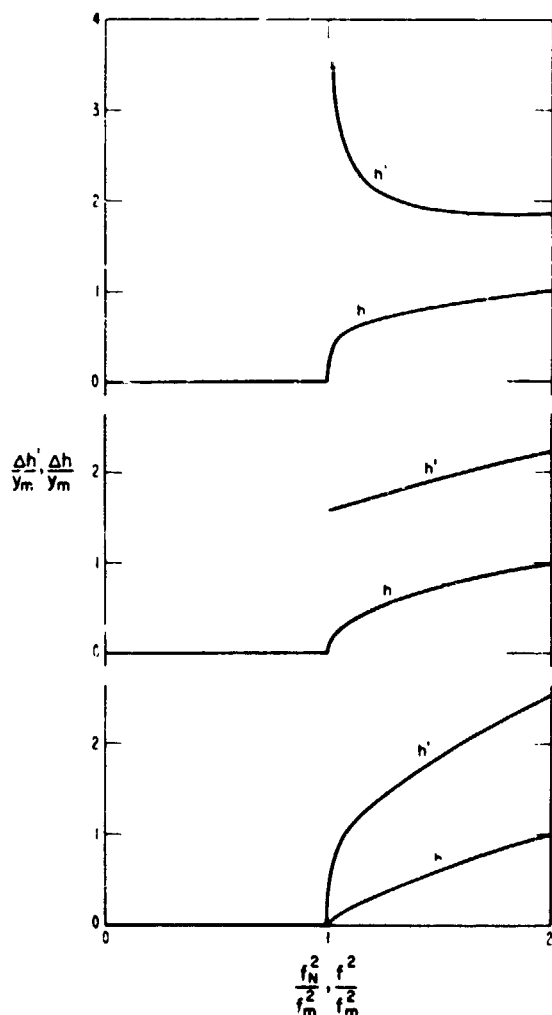


Figure 3.— Model profiles given by 2.10 and corresponding virtual heights.

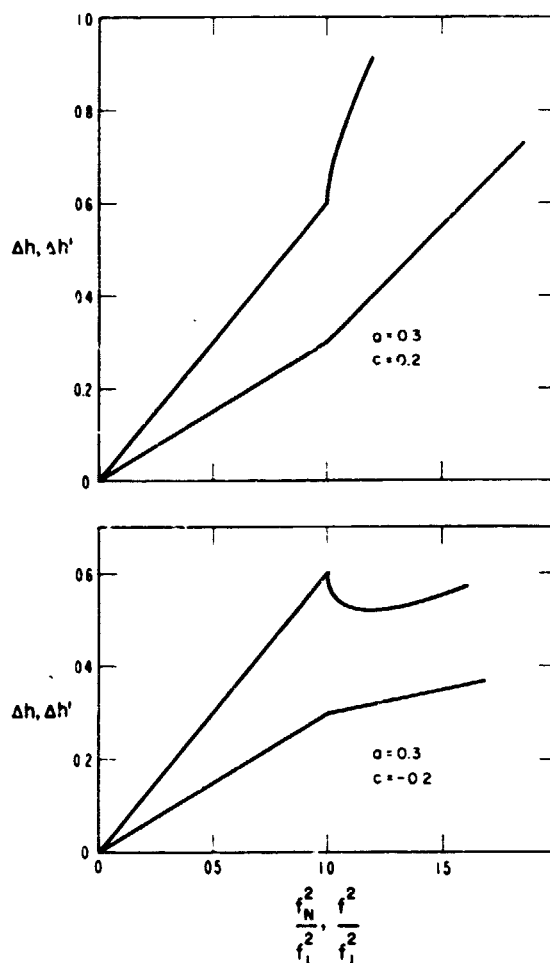


Figure 4.— Profile (h) and virtual-height curve (h') for a model with a discontinuous slope.

slope is increased for $f_N^2 > f_1^2$ in the lower portion the slope is decreased. The result is an infinite slope of the h' curve, with h' increasing vertically in the upper model and decreasing vertically in the lower one.

Cusps like the one shown in the lower part of figure 4 are often observed in ionograms. On the other hand, if the inversion of the integral equation is based on a lamination method, the model used should permit a continuous slope everywhere except where a cusp indicates a discontinuity or at least a very fast change of the slope; otherwise, the h' curve recalculated from the result will fit the original virtual heights very poorly. In other words, at least a second-order approximation should be used within each segment to allow for continuous slopes over the whole profile to assure satisfactory convergence to the true shape of the profile. Furthermore, it can be shown that a finite discontinuity of the second derivative of the profile will always result in a continuous virtual heights and continuous slopes of h' , which means that in most cases a second-order model will also be sufficient for good convergence.

A SECOND-ORDER INVERSION PROCESS

A second-order method can easily be derived by performing a partial integration of (3):

$$\begin{aligned} h'(f) &= \int_0^{f^2} \frac{1}{\sqrt{1-X}} z' df_N^2 \\ &= \left(z' \int_0^{f_N^2} \frac{1}{\sqrt{1-X}} df_N^2 \right)_0^{f^2} - \int_0^{f^2} z'' \int_0^{f_N^2} \frac{1}{\sqrt{1-X}} df_N^2 df_N^2 \end{aligned} \quad (11)$$

We now divide the integration range in several intervals and assume that z'' is constant within each interval while z' is continuous also at the interval limits. If we now define

$$g(f_N^2) = \int_0^{f_N^2} \frac{1}{\sqrt{1-X}} df_N^2 \quad (12)$$

and $g_i = g(f_i^2)$, we can write

$$h'_n = \sum_{i=1}^n (z'_i g_i - z'_{i-1} g_{i-1}) - \sum_{i=1}^n z''_i \int_{f_{i-1}^2}^{f_i^2} g(f_N^2) df_N^2 \quad (13)$$

With the assumption of a constant second derivative, we have

$$z''_i = \frac{z'_i - z'_{i-1}}{f_i^2 - f_{i-1}^2} \quad (14)$$

If all the z'_i for $0 \leq i \leq n-1$ are known, Eq. (13) contains only one unknown z'_n for which this equation can be solved. Therefore, if the height and the first derivative for $f_N^2 = 0$ are known, the integral equation can be inverted

step by step. We first obtain the derivatives at the interval limits. The second derivative is then obtained by Eq. (14), and finally integration over the first derivatives gives the height

$$h_n = h_{n-1} + \frac{1}{2}(z'_n + z'_{n-1})(f_N^2 - f_{n-1}^2) \quad (15)$$

The same principle can obviously also be used in the general case if the earth's magnetic field is included. We only have to replace $1/\sqrt{1-X}$ by μ' of the component used.

A computer program based on this principle has been used very successfully in the Boulder Laboratories for almost 10 years for inversion of ionograms to electron density profiles of the ionosphere. The method has advantages in addition to the high accuracy obtainable. The assumptions made can be used consistently for interpolation, ray tracing, and extrapolation for the maximum of a layer, since z' and z'' are obtained explicitly. There are, however, two basic difficulties in the calculation of electron density profiles. First, we can never obtain h' for very low frequencies because of the so-called starting problem; second, the ionosphere consists often of several layers with minima of electron density in between, and we observe discontinuities or irregularities of h' on the ionograms. As demonstrated earlier, there is no way to obtain a unique solution if only h' of one component is used in the inversion process. The joint use of both components can help to reduce the range of uncertainty.

JOINT USE OF BOTH COMPONENTS

Since the two magnetoionic components, ordinary and extraordinary ray, have different reflection conditions we have to compare them for corresponding radio frequencies, so that both are reflected at the same height. If f_H is the electron gyrofrequency of the earth's magnetic field, such a pair of frequencies f_O and f_X are related by

$$f_X = \frac{1}{2}f_H + \sqrt{f_C^2 + \frac{1}{4}f_H^2} \quad (16)$$

In general, for any given height between the lower boundary of the ionosphere and the reflection height, the group refractive indices are different for those two frequencies; therefore, we expect to improve the results, where uncertainties are involved, by the joint use of both components. The expression for the virtual heights of a monotonous layer for the two components are

$$h'_O = h_O + \int_0^{f^2} \mu'_O z' df_N^2 \quad (17)$$

$$h'_X = h_O + \int_0^{f^2 - ff_H} \mu'_X z' df_N^2 \quad (18)$$

It is convenient to change our notation by introducing the plasma frequency at the reflection level by

$$\phi = \begin{cases} f & \text{ordinary component} \\ \sqrt{f^2 - ff_H} & \text{extraordinary component} \end{cases} \quad (19)$$

and then

$$X = \frac{fN^2}{\phi^2} \quad (20)$$

Both group refractive indices now have the same singularity and can be written as

$$\mu'_O = \frac{M_O}{\sqrt{1-X}} \quad \text{ordinary component} \quad (21)$$

$$\mu'_X = \frac{M_X}{\sqrt{1-X}} \quad \text{extraordinary component} \quad (22)$$

where M_O and M_X are slowly varying functions of X and ϕ .

Abel gave an exact solution for the type of integral equation we have in Eq. (3). For $M_O = 1$ (at the magnetic equator) it is given by

$$h(\phi_n) = \frac{2}{\pi} \int_0^{\phi_n} \frac{1}{\sqrt{\phi_n^2 - \phi^2}} h'(\phi) d\phi \quad (23)$$

We shall attempt to find a solution of the general integral Eqs. (17) and (18) by modifying Abel's solution to

$$h(\phi_n) = \frac{2}{\pi} \int_0^{\phi_n} \frac{S_O}{\sqrt{\phi_n^2 - \phi^2}} h'_O(\phi) d\phi \quad (24)$$

and

$$h(\phi_n) = \frac{2}{\pi} \int_0^{\phi_n} \frac{S_X}{\sqrt{\phi_n^2 - \phi^2}} h'_X(\phi) d\phi \quad (25)$$

for ordinary and extraordinary components, respectively. The functions $S_O(\phi, \phi_n)$ and $S_X(\phi, \phi_n)$ can be determined numerically without difficulties. An example is given in figure 5.

As mentioned earlier, Eqs. (24) and (25) are only correct if the electron density is not decreasing with increasing height; otherwise, we will have different values for the two integrals in (24) and (25), respectively. If we formally apply this solution method and obtain different values, then this difference is due to a discontinuity of the reflection height if a singularity or discontinuity of h' is involved, or it has to be attributed to low electron densities for which no observations were obtainable. If we find such a difference we know that there is a "valley" between two layers

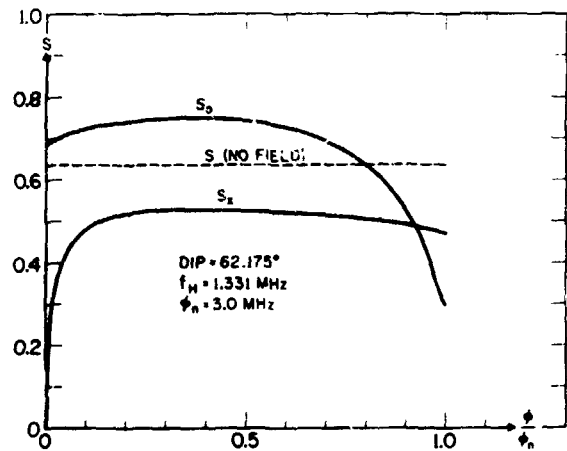


Figure 5.— Functions $S_O(\phi, \phi_n)$ and $S_X(\phi, \phi_n)$, solid lines. The corresponding function without magnetic field equals the constant $2/\pi$ shown as dashed line for comparison.

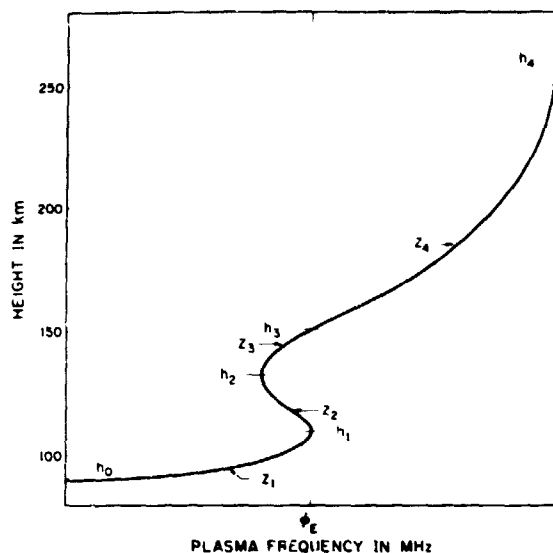


Figure 6. - Model ionosphere with valley.

or there is a significant amount of ionization at low densities, but the question now is whether the profile of those parts can be calculated from this difference which depends on frequency.

Without going into the mathematical details we will present the principles for studying this question. We consider an electron density distribution consisting of two layers as shown in figure 6. We apply formally the solution process to ordinary and extraordinary components at corresponding frequencies and take the difference of the results. We obtain a new integral equation for the effective slope z'_{eff} of the unobserved portion of the profile.

$$D(\phi_n) = \int_{\phi_{min}^2}^{\phi_{max}^2} z'_{eff} K(f_N, \phi_n) df_N^2 \quad (26)$$

where D is the difference of the results and K is the new kernel function, which can be determined numerically. The effective slope is the sum of the absolute values of the two (or more) slopes of the actual profile at each plasma frequency in a valley between the two layers. The integral over the effective slope gives the correct height increment for example $h_3 - h_1$ in figure 6. A few examples of the kernel function are shown in figure 7. The kernel function depends strongly on the dip angle and weakly on frequency. A first-order estimate for the height step or "valley width" can be obtained in the following way. Since z'_{eff} is always positive we can apply the mean value theorem for integrals

$$D(\phi_n) = K_i \int_{\phi_{min}^2}^{\phi_{max}^2} z'_{eff} df_N^2 = K_i(h_3 - h_1) \quad (27)$$

where K_i is a value of the kernel function somewhere in the interval between ϕ_{min} and ϕ_{max} . The lowest plasma frequency in the valley ϕ_{min} is unknown. So we see that for small dip angles, if ϕ_{min} is not very close to ϕ_{max} , the kernel function is almost constant and relatively good estimate for K_i and therefore for the height step can be obtained. This is not true for dip angles around 40° since the kernel function is then very small. For larger dip angles, if the valley is not too shallow and not too deep, again a fairly good estimate of the height step is possible. The uncertainty of K_i increases again towards 90° dip angle, and therefore the valley width becomes less accurate.

We may conclude that depending on dip angle a first-order estimate of the height step may be possible, but we also realize that no further details of the shape of the valley are obtainable, at least in practice, where the accuracy of the virtual heights is limited.

ERROR ESTIMATES AND FIRST ORDER CORRECTIONS

We have seen that the accuracy of an estimated height step depends mainly on the dip angle if both components are used in the inversion process. In practice, there is no chance to learn much more about the distribution in a valley or in the low-density portion of the profile. This is partly due to the limited accuracy of the virtual heights, but in

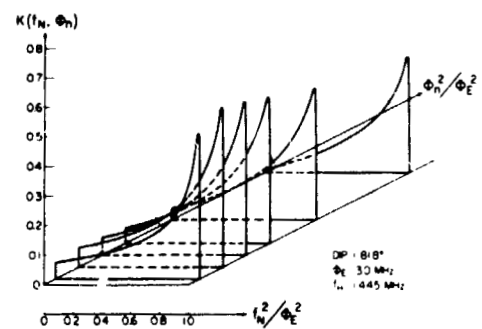
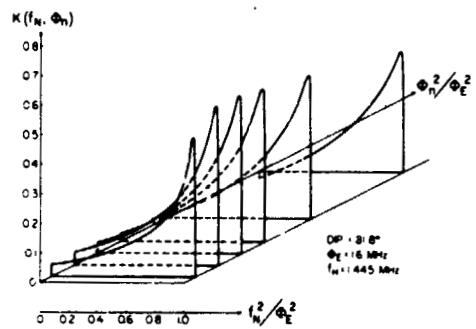
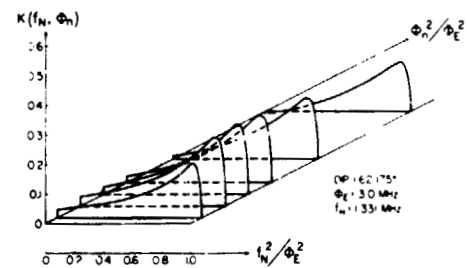
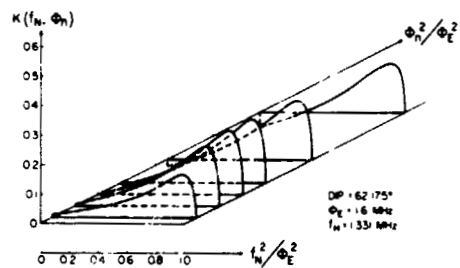
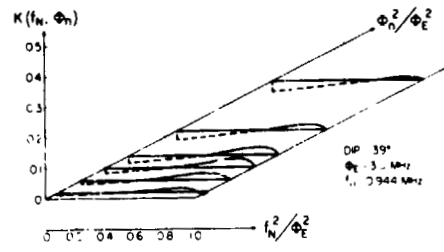
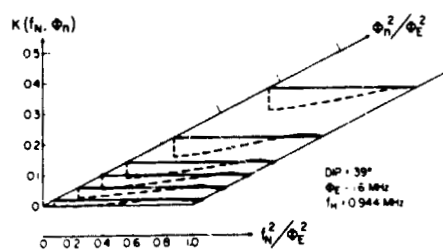
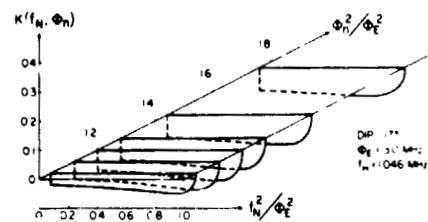
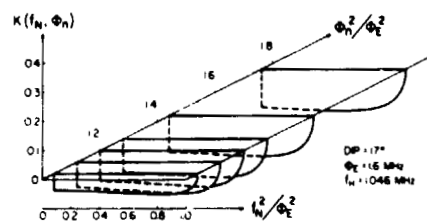


Figure 7.— Kernel function $K(f_N, \phi_N)$ for starting and valley problems at several latitudes. Cross sections $\phi_N = \text{const.}$ through the three-dimensional surface show the dependence of K on f_N^2 . The variation of K with magnetic dip may be seen by comparing the four figures in a column; the variation with ϕ_E for constant dip, by comparing the two figures in the same row.

many cases even more to the inconsistency of ordinary and extraordinary components if the ionosphere is not exactly horizontally stratified. Even if such a height step can be obtained accurately the retardation of the signal within this height interval has to be known for an exact calculation of the remainder of the profile. On the other hand, it is in the nature of the inversion process that errors are reduced relatively fast with increasing frequency. This means, for example, that for a first-order estimate of the "width" of a valley between E and F regions and a reasonable model for electron density distribution in the valley, the error in height for densities near the maximum of the F region will be in the order of 1 km, even if the height step for the valley is wrong by 10 km, provided that the maximum density of the F region is at least 3 or 4 times the maximum density in the E region.

More insight in the error-reducing process can be obtained in the following way: We assume the profile is calculated up to a height $h_c(\phi_{n-1})$, the index c indicating that it is only approximately correct. The virtual height $h'(\phi_n)$ is now used to calculate the portion of the profile between ϕ_{n-1} and ϕ_n . The resulting height $h(\phi_n)$ is again incorrect. If an index e indicates the exact profile, we have

$$h'(\phi_n) = h_0 + \int_0^{\phi_n^2} \mu' z'_e df_N^2 \quad (28)$$

On the other hand, we use the equation

$$h'(\phi_n) = h_0 + \int_0^{\phi_{n-1}^2} \mu' z'_c df_N^2 + \int_{\phi_{n-1}^2}^{\phi_n^2} \mu' z'_c df_N^2 \quad (29)$$

to determine the height interval between ϕ_{n-1} and ϕ_n . If we take the difference between Eqs. (28) and (29), we obtain

$$\int_0^{\phi_{n-1}^2} \mu' (z'_e - z'_c) df_N^2 + \int_{\phi_{n-1}^2}^{\phi_n^2} \mu' (z'_e - z'_c) df_N^2 = 0 \quad (30)$$

If we now assume that the difference between the correct and incorrect slopes is always positive or always negative, the mean value theorem can be applied and we obtain

$$\mu'_l \left[h_e(\phi_{n-1}) - h_c(\phi_{n-1}) \right] + \mu'_u \left[h_e(\phi_n) - h_e(\phi_{n-1}) - h_c(\phi_n) + h_c(\phi_{n-1}) \right] = 0 \quad (31)$$

or

$$h_e(\phi_n) - h_c(\phi_n) = \left(1 - \frac{\mu'_l}{\mu'_u} \right) \left[h_e(\phi_{n-1}) - h_c(\phi_{n-1}) \right] \quad (32)$$

where μ'_l is some value of μ' in the lower interval $0 \leq \phi \leq \phi_{n-1}$ and μ'_u correspondingly in the upper interval $\phi_{n-1} \leq \phi \leq \phi_n$. From Eqs. (21) and (22) we know that $\mu'_l \leq \mu'_u$ and since $\mu' \geq 1$ we see that the height error at ϕ_n is less than the error at ϕ_{n-1} .

$$\Delta h(\phi_n) < \Delta h(\phi_{n-1}) \quad (33)$$

since

$$0 < 1 - \frac{\mu'_0}{\mu'_u} < 1 \quad (34)$$

This property of the inversion process is very important, especially if limited computing facilities exclude the joint use of both components for electron density calculations. In this case reasonable assumptions must be made about the undetected portions of the profiles. While the use of both components permits an error estimate based on a least-squares fit of the recalculated virtual heights to the observed virtual heights of both components, no error estimate can be obtained if one component is used together with some assumptions. On the other hand, since we know that the errors are decreasing with increasing frequency, we can expect that the profile calculated will come close to the true profile at least near a maximum of a layer, often the most important portion — for example, for a world-wide study of the ionosphere — or its time behavior at a fixed location.

It is standard practice to extrapolate h' linearly to zero frequency for the starting process if the slope of the virtual height is positive at the beginning of the ionogram. A valley correction of about the same sophistication can be applied easily using the following rule for the input data h'_i if h'_{obs} are the observed data

$$h'_i = h'_{obs} \quad f \leq f_c E \quad \text{and} \quad \phi \leq f_o E \quad (35)$$

$$h'_i = 2h'_{obs} - h'_{min} E \quad f \leq f_c E \quad \text{and} \quad \phi > f_o E \quad (36)$$

$$h'_i = h'_{obs} \quad f > f_c E \quad \text{and} \quad \phi > f_o E \quad (37)$$

which simply means that for the calculation of the F layer we double the virtual path within the E layer corresponding to the assumption that the E layer is symmetrical, which can be roughly justified by the physics of layer formation in the ionosphere.

A similar rule can be applied to nighttime ionograms if the virtual heights are decreasing with increasing frequency at the beginning of an ionogram. In this case, we know that there is another layer below the observed one with a critical frequency less than the frequency where the first echo is seen. Sometimes it is possible to extrapolate for this critical frequency, and the retardation in this lower layer can be estimated by simple assumptions such as a constant slope in the F region and a parabolic profile of the E region. If such an extrapolation of the critical frequency is not practical we may instead use

$$\begin{aligned} h'_i &= 2h'_m - h'_{obs} & f < f_{min} \\ h'_i &= h'_i & f > f_{min} \end{aligned} \quad (38)$$

where h'_m is the minimum virtual height at the frequency f_m . Equation (38) surely is harder to justify than Eqs. (35) through (37), but it assures at least that the correction is in the right direction.

Summarizing we see that the accuracy of profile calculations is always limited. Sometimes it is not even possible to obtain a reliable error estimate in the sense that the errors of the results should bear a simple relation to the estimated errors of the virtual heights. This holds at least for the valley and starting problem. There, depending

mainly on the dip angle, the joint use of both components can help to reduce the errors significantly. The error is decreasing with increasing frequency, but it also depends very much on the structure of the ionosphere. Nevertheless, in many cases the results in the vicinity of a maximum of a layer are more accurate than the virtual heights, due to the error reduction inherent in the process as can be demonstrated by model studies.

DISCUSSION

Reinisch: Is the programming required for routine work already done? If so do you now use it?

Paul: This is just a study and not in routine use.

Wright: You are asking if this Abel generalization has been used in a working method to deduce valleys. The $N(h)$ formulation that Dr. Paul has devised is in use to obtain profiles. Another process that makes use of the ordinary and extraordinary information is in routine use to resolve the valley ambiguities so far as permits. To perform this solution, some assumptions are needed about the valley parameters.

Reinisch: Do you have any figures on the frequency of occurrence of the valleys and its average depth?

Paul: We cannot detect that. You can do something like Jackson did. Your results depend on the model.

AUTOMATIC $N(h, f)$ PROFILES OF THE IONOSPHERE

WITH A DIGITAL IONOSONDE

J. W. Wright, A. R. Laird, E. J. Violette

D. McKinnis, and D. Obitts

National Oceanic and Atmospheric Administration, Boulder, Colorado

N73-11606

ABSTRACT

A method is described to accomplish automatic data selection and profile inversion to obtain ionospheric electron density profiles from digitized radio soundings. The profile inversion is based on the well-established formulation of Paul [1960] by which the optimum radio frequency sounding intervals can be specified from an approximate knowledge of the profile; the expected virtual height coordinates (h') at these frequencies (f) are estimated, and procedures are then used to select $h'(f)$ observations nearest the predicted coordinates from a subsequent digital ionogram. From these the next profile is obtained. The process adaptively follows the changing shape and detail of the profile. The procedure requires an average of 15 sec per profile on a standard data processing computer, and can be adapted, with benefit to online real-time use in our digital ionosonde.

INTRODUCTION

Long before the age of digital computers, the panoramic ionosonde had established the swept-frequency ionogram as a convenient and subjectively informative representation of ionospheric structure. To a considerable extent phenomena in the ionosphere are recognized, described, and named – if not necessarily understood – by their ionogram manifestations. These qualitative and pictorial applications sufficiently justify systems designed to make such recordings, but the greater usefulness of quantitative information has led to progressively more elaborate schemes for their analysis. Radio echoes from virtually the entire near side of the ionosphere are recorded on the program; the inversion problem, and most of the formidable difficulties of deducing the profile, has been recognized from the beginning. But in part because it was relatively easy to “scale” simple parameters in great quantity from ionograms, and also because profile inversion involves awesome arithmetic, only scattered efforts to deduce profiles were attempted until electronic computers emerged to assume the principal burden. Even so, 20 to 50 virtual height-frequency readings are required as input information to the inversion process; when this scaling is done by hand from film ionogram recordings, the labor remains excessive and accuracy compromised; yet the task requires experienced judgment, since much of the echo information of the ionograms is irrelevant to the profile.

The next stage in the evolution of these techniques is currently under way: the application of digital hardware, which we believe provides an approach to the solution of the remaining problems. Small, fast, general purpose digital computers are now available, at a fraction of the cost of inflexible special-purpose, fixed-wire control systems. When interfaced to the essential elements of the ionosonde (transmitter, receiver, frequency-synthesizer) the general-purpose computers assume all of the functions of system control, echo recognition, parameter digitization, and preliminary or even final data processing. This is the orientation of our program of digital ionosonde (Dynasonde) development [Wright, 1969], and it consists of three interdependent areas of activity:

1. Development of the Dynasonde system itself, including the selection of digital and analog system components, interfacing, and the refinement of control software for the Dynasonde online general-purpose computer.

2. Development of the Kinesonde, an independent and more narrowly oriented hardware system in which we concentrate on the measurement and statistical analysis of ionospheric radio echo parameters as time series, for studies of ionospheric motions and microstructure; these additional measurement capabilities are then achievable directly in the Dynasonde system through straightforward software design.
3. Development of data analysis methods to accomplish specific measurement objectives such as echo location (angle of arrival) and ionospheric "drifts" from spaced-antenna measurements, or the extraction of ionospheric electron density profiles from digitized group-path versus frequency measurements. Although our goal in each case is to accomplish these analyses online and in realtime (within the Dynasonde), it is expedient to approach their development offline, using tape recorded digital data and a large, general-purpose, easily programmed central computer.

The paper reports our progress toward the specific objective of determining ionospheric electron density profiles automatically -- that is, without the need for manual data scaling and subjective echo pattern recognition. Measurements for this purpose are being made by the prototype Dynasonde, the main features of which have been described recently [Wright, 1969], and are not repeated here, except to illustrate pictorially the digital ionograms themselves. Our approach to automatic profiles is heavily influenced by Paul's [1960] illuminating formulation of the profile inversion problem and developed into our present general-purpose working system by Paul [1967], Howe and McKinnis [1967], and Wright [1967]. In the section we outline this formulation, emphasizing some of its properties especially appropriate to our present objective. We also describe the "AutoN(h)" process, which consists of a group of computer subroutines each of clearly definable structure, purpose, and interrelationship. Some results of the AutoN(h) system are illustrated.

FORMULATION OF THE PROFILE INVERSION PROBLEM

The vertical component of the group path of a vertically incident radio wave pulse in a plane-stratified ionosphere determines one observable parameter of the pulse -- its total travel time. One-half of this, multiplied by the speed of light, gives the apparent or "virtual" height h' for the center frequency f of the radio wave pulse. Thus, the group speed varies along the path z to the height of reflection h , and the resulting virtual height varies with probing frequency as

$$h'(f) = \int_0^h \mu'(f, z) dz \quad (1)$$

Here the group refractive index μ' is the ratio of the speed of light to the group speed, and depends on the frequency of the radio wave, the direction θ of propagation with respect to the earth's magnetic field, and the properties of the ionospheric medium, electron density and magnetic field strength, expressed through the plasma frequency f_N and electron gyrofrequency f_H . In our problem f_H and θ are constant for a given location, and f_N is assumed to vary with height (monotonically) and time only. Although μ' is double valued, resulting in two values of f_N for total reflection of f (identified as f_o and f_x , respectively), we discuss mainly the analysis of $h'(f_o)$ to obtain $z(f_N)$. The observations also provide $h'(f_x)$, which may be analyzed independently to give $z(f_N)$; in principle, the results should not depend on the choice of o or x , but in practice the assumptions of monotonicity and of only horizontal stratification (and sometimes profile errors) cause inconsistencies when $h'(f_o)$ and $h'(f_x)$ results are compared. In part, the o and x information can be reconciled to advantage in allowing a first order correction for nonmonotonicity in the profile [Paul and Wright, 1963; Howe and McKinnis, 1967], but the AutoN(h) development defers this problem and employs $h'(f_o)$ alone.

The inversion of Eq. (1) to obtain $z(f_N)$ cannot be expressed in closed form because of the complexity of μ' . Many step-by-step formulations have been advocated [reviewed by Wright and Smith, 1967], but only Paul [1960] explicitly applies the basic mathematical-physical requirements of the problem to justify a workable numerical process. Approaching reflection, where $f_N(z) = f_0$, μ' tends to infinity, so that a step-by-step process for the inversion of Eq. (1) based on slabs of unknown thickness Δz is impossible. The unknown height is therefore removed from the integral limit, and the integration is obtained over the independent variable f_N , or any monotonic differentiable function of it $\phi(f_N)$, by the transformation

$$h'(f) = h_0 + \int_{\phi_0}^{\phi(f)} \mu'(f, f_N) \frac{dz(f_N)}{d\phi(f_N)} d\phi(f_N) \quad (2)$$

where h_0 and ϕ_0 refer to the bottom of the ionosphere.

To simplify the notation, and to represent the true height and the retardation separately, we may rewrite Eq. (2) as

$$h' = h + \int_{\phi_0}^{\phi(f)} (\mu' - 1) z' d\phi \quad (3)$$

Many methods of inversion now proceed by assuming that $z'_i = (dz/d\phi)_i = \text{const}$ over short (not necessarily uniform) intervals of ϕ , whereupon the integral in Eq. (3) can be replaced by a summation and the resulting system of equations solved for the several unknowns z'_i , given an equal number of h'_i . There are two defects of this procedure, however. The weight function μ' remains infinite at the reflection level, so that the usual assumption — that the constant z'_i represent mean values attributable to the middle of their corresponding intervals — is systematically wrong. This defect might be ameliorated, as Paul [1960, 1971] has proposed, by specifically assuming that the value of z' obtained refers to the reflection level, rather than to the interval midpoint. However, a more basic problem exists since ionograms show us very clearly that z' is *not* discontinuous, except perhaps approximately so at a small number of well-marked "cusps" of the $h(f)$ observations. To conform with the observations, it is essential that the continuity of z' be preserved by the inversion process except where the observations indicate discontinuity. If Eq. (3) is integrated by parts and z' is assumed differentiable, we obtain

$$h'(f) = h(f) + \left[z'(f) \int_{\phi_0}^{\phi(f_N)} [\mu'(f, f_1) - 1] d\phi(f_1) \right]_{\phi_0}^{\phi(f)} - \int_{\phi_0}^{\phi(f)} dz(f_N) z''(f_N) \int_{\phi_0}^{\phi(f_N)} d\phi(f_1) [\mu'(f, f_1) - 1] \quad (4)$$

Note that the retardation is now expressed by two terms in which the (variable) first and second derivatives along the profile are weighted by a quantity $(\mu' - 1)d\phi$, which is everywhere finite. If the integration of the term involving $z'' = d^2z/d\phi^2$ is now replaced by a sum over i intervals in each of which z'_i is assumed constant, we obtain a system of equations relating discrete values of the observable quantity h'_i to discrete values of the continuous profile derivatives z'_i at the reflection heights h_i . In our application of these ideas, we take $\phi = \log f_N^2$, and define the integration variable $t = \log f_N^2 - \log f^2$. The system of equations relating h'_i and the profile may then be written

$$h'_i = h_i + \sum_{j=1}^i [z'_{j-1} g_{ji} + z''_{j-1} (g_{ji} - g_{j-1, j})] \quad (5)$$

where

$$g_{ji} = \int_{t_{j-1}}^{t_j} (\mu' - 1) dt \quad G_{ji} = \int_{t_{j-1}}^{t_j} (\mu' - 1) t dt \quad (6)$$

The quantities g and G depend on the plasma frequencies used to divide the profile and on the radio frequency, but not on the profile itself. A simple table of g and G would suffice for all profiles if the intervals were to be specified arbitrarily and kept invariable. However, it must be remembered that Eq. (5) is an "exact" relation between the profile and h' provided the z_i'' are constant over successive intervals. The relation is thus a good approximation if the change in z'' is sufficiently small between successive intervals; and as the form of the profile varies from one occasion to the next, the optimum interval widths (hence the optimum measurement frequencies for h') will likewise vary.

The essential principles of our strategy for automatic $N(h)$ calculations can now be stated as follows:

1. Corresponding to a given change of z'' along the profile, there is a maximum permissible interval width $\Delta \log f_N^2$ within which z'' is approximated by a constant value, such that the height error Δz does not exceed a permitted maximum. In other words, the profile, once known, can be partitioned along the scale of plasma frequencies to determine optimum intervals at which radio frequencies should be chosen to obtain virtual heights from which to determine the profile.
2. The apparently circular argument of (1) becomes useful when two other principles are realized:
 - (a) Optimum intervals may be determined in an iterative fashion if an approximate profile is known or can be assumed.
 - (b) In a sufficiently continuous time-series of ionograms, the profile from one ionogram serves as an approximation from which to determine the optimum intervals for measurement of the next ionogram.

THE AUTON(h) PROCESS

The system consists of a group of subroutines programmed (at present) in FORTRAN IV for use on a large central computer. The observational data are tape-recorded sequences of digital ionograms obtained by the Dynasonde. The ultimate aim of adapting the method to online real-time use within the Dynasonde has been kept in view, and is commented on later. Figure 1 is a block diagram giving the principal subroutine names and interrelationships. Let us break into the loop following a PROFILE calculation and trace the subsequent logical operations:

1. Subroutine GETF uses the result of the PROFILE calculation to determine optimum sounding frequency intervals such that the height interval error caused by assuming $z'' = \text{const}$ in each frequency interval will not exceed a certain chosen limit.

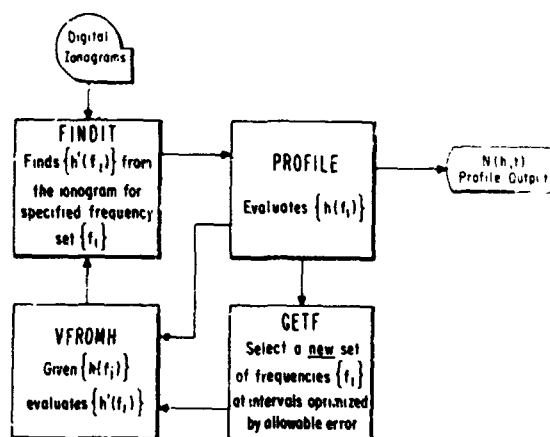


Figure 1.— Block diagram of the principal subroutines of the AutoN(h) system showing briefly their functions and interrelationships.

2. For the newly specified set of sounding frequencies, and using the given true-height profile, subroutine VFROMH calculates the expected virtual heights. This is done by a forward calculation using Eq. (5).
3. Equipped now with optimized frequencies and their expected virtual height, FINDIT examines the next digital ionogram in the sequence at these $h'(f)$ coordinates to select accurate values of the nearest $h'(f)$ data.
4. These values are transferred to the PROFILE subroutine, which solves Eq. (5) in the backward direction for the wanted profile.

The operation of these subroutines is of course somewhat more complicated than this simple description might suggest. PROFILE and VFROMH involve basically deterministic arithmetic treatments of Eq. (5) in opposite "directions", and each, given the other's output, would produce the other's input identically.

Similarly, GETF is deterministic: If we consider the Taylor series expansion of the profile at a given point t_j and for $z'(t_j)$ write z_j' :

$$h(t) = h_j + z_j'(t - t_j) + \frac{1}{2} z_j''(t - t_j)^2 + \dots \quad t_j \leq t \leq t_{j+1} \quad (7)$$

The profile is found using $z'' = \text{const}$ in each interval, but this yields a succession of values of z'' that can be used to estimate the amount by which z'' is not constant along the profile. Thus, the difference in z'' between successive intervals gives an estimate of z_j'''

$$z_j''' = 2 \left| \frac{z_j'' - z_{j-1}''}{t_{j+1} - t_{j-1}} \right| \quad (8)$$

as well as an estimate of the error resulting from the use of constant z'' .

If the most important part of the error arises in the first term dropped from the power series, it is

$$\epsilon \simeq \frac{1}{6} z_j'''(t - t_j)^3$$

Thus an allowed frequency step can be chosen proportional to $(\epsilon/z''')^{1/3}$ to generate an optimized set of frequencies.

To a limited extent, the error permitted in each interval automatically controls both the number of intervals used to divide the profile and the resulting accuracy of the profile, and this "control" would probably work smoothly and monotonically with the permitted error if the ionogram (and the profile itself) were "smooth". In fact, however, a wide spectrum of profile fluctuations exists; it might be feared that a small permitted error would lead to a "runaway" condition in which small detail of the $h'(f)$ curve — or the digitization noise itself — would demand its own cluster of closely spaced intervals. Fortunately, several factors work against this tendency. One, under the user's control, is a frequency-height window to be described, which smooths over $h'(f)$ detail of the order of its own size. Perhaps more important, the GETF process has little tendency to add detail that is not already present in the preceding profile information. In this sense, the permitted error must be interpreted as an average value, appropriate to a smooth curve over the interval, whatever detailed variations take place within it.

The selective effect of a given permitted error in determining the degree of resolution of the ionogram and profile is illustrated in figure 2, which shows the dependence (for a particular sequence of ionograms) of the number of

"scaling" intervals on alternate choices of the permitted error ϵ . Results for two values of ϵ are shown, and for each ϵ , two initial sets of scaling intervals were chosen, comprising 5 and 34 intervals, respectively. As shown, about 10 intervals suffice for the ionogram sequence if $\epsilon = 50$, and about 25 intervals are required if $\epsilon = 10$. The number of intervals tends toward these partitioning values, whatever the number of intervals in the initial set. This also illustrates that the $\text{Auto } N(h)$ process can adaptively follow smooth, natural changes in the amount of structure of the profile, and that the accuracy of the automatic process is not necessarily constrained by the amount of detail included in the initial scaling.

Figure 3 represents a digital ionogram from the Dynasonde tape recordings graphed by a computer-drive microfilm plotter with scale proportions chosen to simulate a conventional analog ionogram. In the upper part of figure 3 two sets of scaling intervals have been plotted on the ionogram. The arrows indicate the frequencies of the sparse initial "scaling", and the O 's are the optimized intervals as calculated in GETF ($\epsilon = 10$) and plotted at the h' coordinates calculated by VFROMH. Because the original scaling does not adequately describe the $h'(f)$ curve, the virtual heights calculated for the optimized frequency set do not fall on the $h'(f)$ curve at all points, particularly near f_oE . Subroutine FINDIT is given the task of reconciling such predicted $h'(f)$ values with the next ionogram of the series.

Subroutine FINDIT is the principal link between the digital ionogram and the profile. It is here that the $\text{Auto } N(h)$ process adapts itself to the time-varying ionosphere, that echoes relevant to the profile are distinguished from those that are not, and that moderate smoothing of the $h'(f)$ data is permitted. FINDIT is provided with a sequence of predicted (h', f) coordinates, and has the task of finding values acceptably near these coordinates from the array of digitized $h'(f)$ data of the ionogram. The initial h' values are usually calculated from the previous profile by VFROMH at frequencies specified from this profile by GETF. Alternatively, FINDIT may be given an approximate set of initial h', f coordinates scaled "by eye"; they could easily be specified without "scaling" at digital display by use of a light pen. These alternatives provide ways of starting the entire process.

Whatever the source of the initial h', f set, FINDIT establishes a "window" of controlled width in $f (= Fw)$ and in $h' (= H'w)$ centered on the given h', f coordinates.

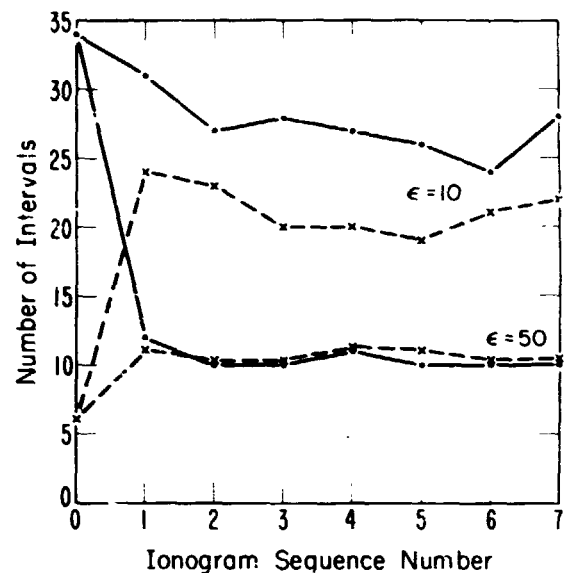
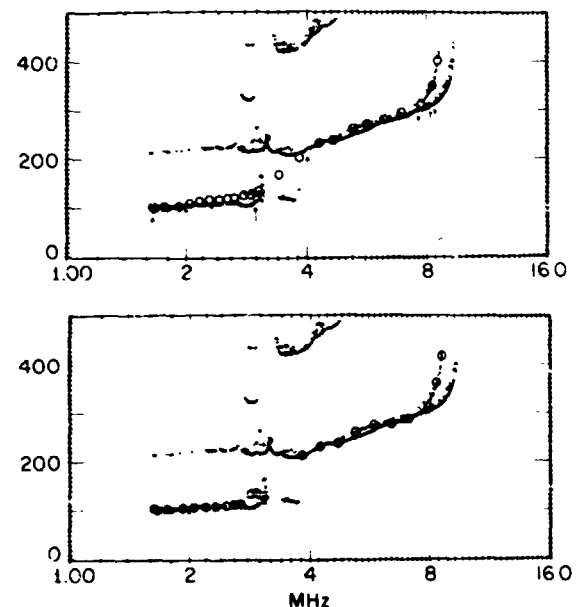
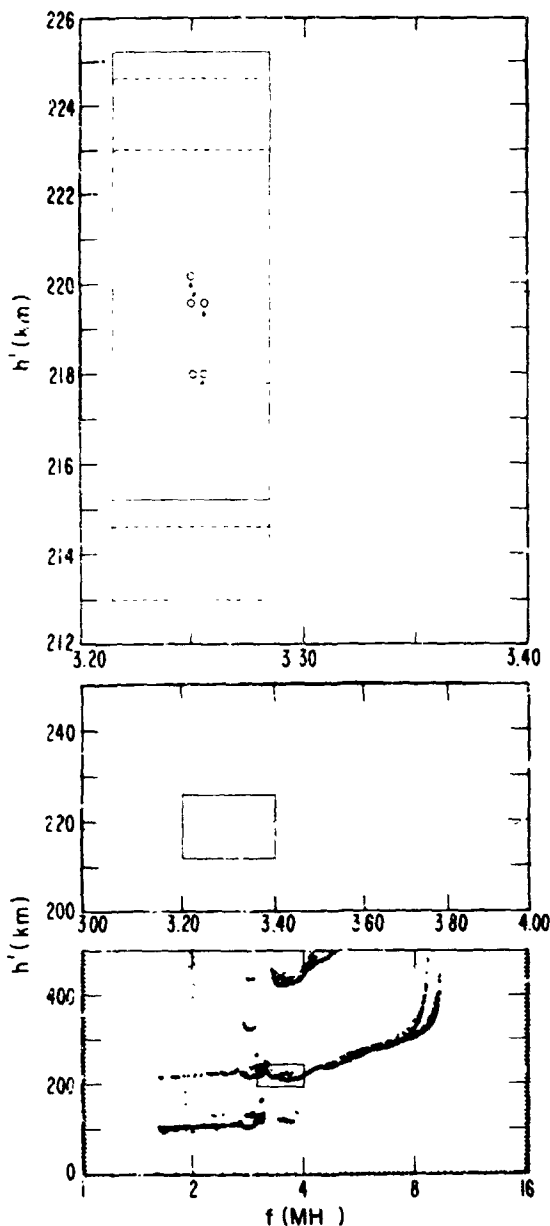


Figure 2.— Number of intervals vs. ionogram sequence number for two values of the error parameter ϵ , showing initial number of intervals at ionogram number 0, and for ionogram numbers 1–7 the number of optimized intervals determined by GETF.



DIGITAL IONOGRAM, BOULDER
2 OCTOBER 1970, 0850:05 MST

Figure 3.— (a) Ionogram number 1 showing the initial sparse "scaling" intervals (arrows) and the optimized intervals (O 's) as calculated by GETF. (b) The echo positions at optimized intervals from GETF, as found by FINDIT, using h' information from the ionogram.



DIGITAL IONOGRAM, BOULDER
2 Oct. 1970, 0850:05 MST

Figure 4.— (a) Full digital ionogram. (b) Area within the box in (a) enlarged ten times in height. (c) Area within the box of (b) again enlarged ten times in height, showing the correction of the estimated h' to the median h' in the window and of f to the frequency coordinate of the median height, in four successive steps. Note that the third and final steps position the window center exactly on measured data heights.

FINDIT first determines the number of h' values within the window, the median h' of these values, their quartile range, and the frequency coordinate of the median h' . If the median h' differs by less than a preset value from the initial value, the median (and its frequency coordinate) are retained as wanted values for the subsequent PROFILE calculation. Otherwise, the height coordinate (only) for the window center is moved to that median value, and a new median and other quantities are found for this new position of the window. This process is repeated a maximum permitted number of times or until the change (in h' only) from one iteration to the next is less than the preset tolerance.

The arbitrary parameters of this process are obvious, and can be chosen within wide limits without significantly affecting the resulting profile. Note that the actual frequency coordinate of the selected echo is reported, but that the window is not permitted to move in the frequency direction. This assures that the frequency intervals remain near (within F_w of) the optimum intervals determined by GETF.

The lower part of figure 3 illustrates the echo selection results of FINDIT, subsequent to the GETF and VFROMH operations that led to the upper part of figure 3. The window motion operations have successfully located the E region echoes near the wanted frequencies.

Figure 4 shows a typical sequence of the window operations. The lower part (a) is again a computer-driven microfilm plot of all of the h', f coordinates digitized by the Dynasonde. The other parts of the figure were also plotted by the computer, and they zoom to successively smaller portions of the $h'(f)$ array. From GETF, an h', f coordinate has been "predicted" at a position within the "box outlined on the computer-plotted full ionogram, and in the middle part of the figure only this region has been replotted, full scale (b). Neither the outlined region in part (a) nor that in part (b) represents the window described above; parts (a) and (b) are intended simply to emphasize the level of ionogram resolution (in h' and f) at which the FINDIT process is operative. In the upper part (c) of the figure the initial center and dashed outline of the window is shown. The 0.2-km digitization resolution of h' and the uniform frequency spacing of the ionograms usually programmed in the Dynasonde are clearly evident at this magnification of the ionogram. The initial $h'(f)$ position does not happen to lie on an echo, but this is rectified in the first step, where the frequency coordinate is set to that of the median h' within the window. The window is

then centered (in h' only) on that median h' , and the process is repeated. In this example, the final window position is shown by the continuous outline.

FINDIT has special provisions for expanding the window should the first look contain fewer than an expected number of echoes; decisively empty windows are simply dropped from further consideration. This is appropriate, for example, for interval limits lying near a decreasing f_oF2 . Conversely, if the highest frequency interval limit seems too low in frequency (according to an empirical criterion that involves an extrapolated critical frequency), then more points are added at the end and their virtual heights are estimated ($V''ROMH$) using the z' and z'' values in the last previously existing interval. Any such "feeler point" that FINDIT can validate by examining the ionogram is retained with its coordinates adjusted to the observed values. Together, these provisions follow increases or decreases of the peak density.

The next step in this loop is the PROFILE calculation using the data obtained by FINDIT from the new ionogram. The PROFILE subroutine is virtually identical to the program described by *Howe and McKinnis [1967]*. Coefficients [Eq. (6)] for each interval are calculated as needed; a parabolic peak is determined by extrapolating $1/z'$, and a topside extrapolation can be performed if desired. The principal description of the profile is in terms of the height and z' values at the several discrete values of $\log_{10} f_N$, or their equivalent values of N_e/cm^3 ; both parameters are tabulated as in table 1. The cumulative total electron content is also evaluated. Since z' is continuous between tabulated intervals, a second-order interpolation can be performed for N_e (or $\log_{10} f_N$) at wanted fixed heights. A subroutine is available which calculates N_e at each 10-km level and provides punched cards with this information.

EXAMPLES OF THE $\text{AutoN}(h)$ PROCESS

As discussed above, the permitted error ϵ and the window size influence the amount of detail the $\text{AutoN}(h)$ process will recognize in a sequence of profile calculations. Figures 5 and 6 are examples in which the process follows fine and gross structure.

Figure 5 shows four E -layer height profiles corresponding to the first, fifth, tenth, and fifteenth profiles in a series spanning approximately an hour. These profiles were calculated with $\epsilon = 2\text{m}$ and involved about 20 intervals to describe the fine detail of the E -region $h'(f)$ information. The initial profile (1002 MST) shows

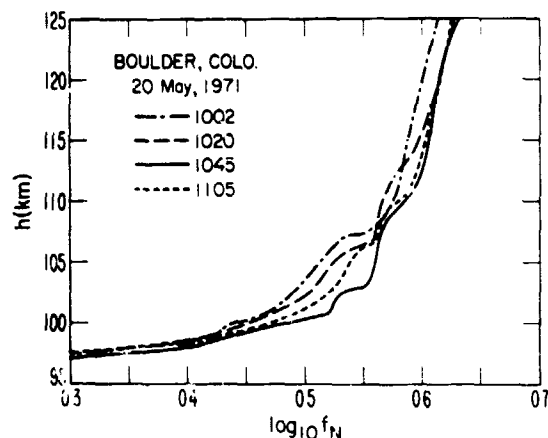


Figure 5.— E -layer $\text{AutoN}(h)$ profiles showing small, but significant height and structure changes over a one hour period.

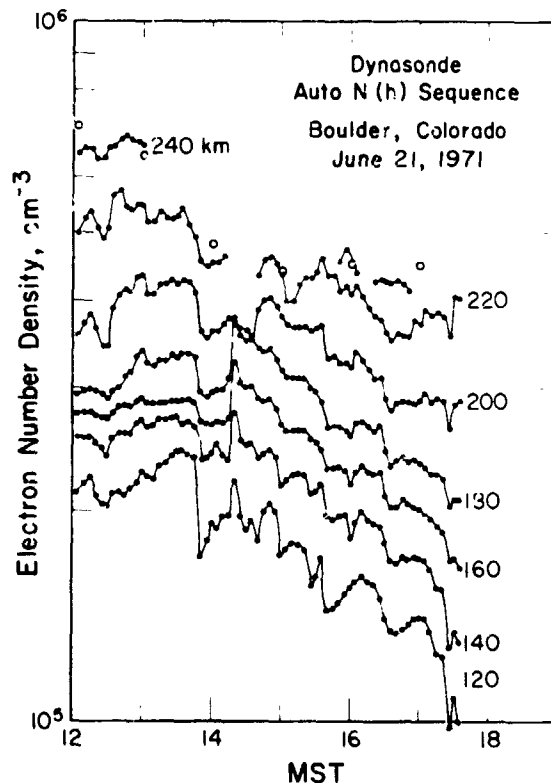


Figure 6.— Electron density variations at fixed altitudes from a 5-1/2-hour period calculated by the $\text{AutoN}(h)$ process from an initial manual scaling at 1205 MST.

TABLE 1.— Tabulated $N(h)$ profile produced by the Auto $N(h)$ process. The first column lists values of $\log_{10} f_N$ representing the optimized frequency intervals found by GETF from the previous profile. The next column (H^*) lists the *observed* h' from the digital ionogram; values of h' are digitized to 0.2 km resolution, and used to this accuracy in the profile calculation, but are rounded in printing here to 1.0 km. The word EXTRAP signifies a provision to extrapolate for underlying ionization, not employed here. Columns H , Z^* and Z^{**} represent the calculated real heights and the first and second height derivatives. The cumulative integrated electron content and the electron density equivalent to $\log f_N$ complete the profile information. The F-layer peak extrapolation is listed on the bottom line. The quantity S represents the quarter-thickness of a parabola continuous (in Z^*) with the profile below.

FOR LOG FX	LOG FN	BOULDER 71 05 20 1036:53				AUTON(H)		
		Q=3. H^*	F=0.560000. H	FH=1.429120. Z^*	SIN 1=0.923900. Z^{**}	INTEGRAL	EL/CC	RESIDUAL
A102	0.1990	EXTRAP	99.40	0.000			3.10118+004	(SD*****)
	0.2990	99.0	99.40	0.000	0.00	0.000	4.91504+004	0.000
	0.3090	99.0	99.40	0.000	0.00	0.000	5.14667+004	0.000
	0.3778	100.0	99.50	1.482	10.77	0.065	7.06500+004	
	0.4749	120.0	103.13	35.838	176.94	3.512	1.10478+005	
	0.5546	199.0	119.40	168.122	829.13	26.089	1.59523+005	
	0.5848	208.0	128.91	147.976	-000334.63	42.343	1.83243+005	
	0.5900	211.0	130.51	155.477	711.96	45.308	1.87743+005	
	0.6066	216.0	135.56	149.640	-000176.28	55.158	2.02618+005	
	0.6239	227.0	141.01	165.174	448.73	66.656	2.19430+005	
	0.6457	267.0	150.05	248.939	1919.05	87.584	2.42631+005	
	0.6653	308.0	161.02	311.871	1608.13	115.486	2.65509+005	
	0.6856	380.0	176.55	452.230	3453.79	158.813	2.91554+005	
	0.6931	367.0	182.46	333.371	-007896.86	176.333	3.01835+005	
	0.7044	341.0	189.05	250.264	-003681.00	196.715	3.17941+005	
	0.7300	325.0	200.49	196.861	-001043.53	235.220	3.57703+005	
	0.7518	324.0	208.80	183.871	-000297.61	266.462	3.95523+005	
	0.7714	334.0	216.24	196.195	314.92	297.250	4.32818+005	
	0.7962	371.0	227.92	274.229	1571.05	350.977	4.85261+005	
	0.8098	401.0	236.15	333.482	2187.07	392.229	5.16497+005	
	0.8165	535.0	244.23	859.764	38850.45	434.717	5.32862+005	
	LOGFM=0.8178		HMAX=249.62	S= 35.334		SHMAX=463.553	5.35852+005	

inflections (corresponding to cusps on the h' curve) at $\log_{10} f_N \sim 0.43$ and ~ 0.53 . About 20 min later the inflection at ~ 0.43 has disappeared, the inflection at ~ 0.53 appears at a lower height, and a new inflection has appeared at ~ 0.57 . The same structure is present at 1045, but at slightly lower heights. At 1105 the profile structure is slightly higher than the previous profile, and the inflections are less pronounced. Careful comparison of the profiles with the ionograms, and with the other profile calculations of the sequence, shows that there are significant, if small, changes of height and structure in the E region. The Auto $N(h)$ process can easily follow evident height variations in density to better than 0.1-km resolution.

Operational conflicts for the Dynasonde make it difficult at present to obtain satisfactorily long sequences of digitized ionograms with which to demonstrate the performance of the Auto $N(h)$ process in following complete diurnal variations. A partial example of this kind, however, is shown in figure 6. The calculated electron density time variations at a number of fixed altitudes on 21 June 1971 are plotted over the period 1205–1735 MST at 5-min intervals. The sequence (of 67 profiles, requiring 15 min computation time) was obtained entirely automatically after initiation by a rough manual scaling of the first (1205) ionogram, which was computer-plotted for this purpose in the format of figure 4(a). The hourly ionograms also were plotted, and the scaled values of $f_o F2$ (converted to electron density) are included on figure 6. After 1730 the $h'(f)$ curve became unrecognizable because of multiple sporadic E reflections. Short-period changes that occur simultaneously and similarly at all altitudes (e.g., at 1540) are probably introduced by the Auto $N(h)$ process itself, by faulty following of the $h'(f)$ information. However, the main features of this part of the diurnal variation are well represented and most of the indicated structural variations are significant.

FUTURE DEVELOPMENTS

The process described here is not in a final form, but it works sufficiently well that it is already of practical utility. The main problems arise when very rapid changes or genuine echo ambiguities occur; often a human analyst would then also not be sufficiently informed by the usual ionogram to resolve the ambiguities. The additional problem with an automatic $N(h)$ process is that, once derailed, it is not certain to get once again on the track of the true profile. In practice, however, our process does realize appreciable stability against such derailments, and the way to further improvement is clear. Even within the $h'(f)$ information alone, the computer can rapidly perform consistency tests – for example, between o and x , among multiple echoes, or between successive ionograms forward and backward in time – that are tedious and relatively ineffective when attempted subjectively.

The Auto $N(h)$ process provides a necessary link between a digitized data-acquisition system (here, the Dynasonde) and the data user. Since the acquisition of fully digitized ionograms (fig. 4) is already an automatic, efficient, and reliable process, the $N(h)$ analysis could be performed centrally from data supplied by remote sounders. If tape recordings are used, the quantity of tape required remains finite since it can be reused. Alternatively, the entire system is compatible with centralized ionosonde control and data processing via any data link of moderate capacity.

The Auto $N(h)$ concepts can provide entirely new capabilities to the Dynasonde when they are included as a part of its own real-time software. Since only a very small fraction of the echo information of an ionogram is required for a profile, and since our process determines both the optimum sounding frequencies and the expected echo arrival times, the usual ionospheric sounding process can be tremendously compressed. While perhaps 2000 real echoes from 1500 discrete frequencies are contained in the average ionogram (requiring 4 times this number of pulse transmissions at 200/sec in our real-time echo recognition process), only about 20 pieces of $h'(f)$ information are required for the profile. When the optimum frequencies are known with their expected echo arrival times, only 1% to 3% of the normal ionogram sounding activity is required. The savings can be realized in many otherwise unattainable advantages, including

1. *Greatly reduced radio interference.* The necessary $h'(f)$ information can be acquired at a random pulse emission schedule, or in a burst of only 0.1 sec duration, and the frequencies are already tangibly "randomized" by the GETF conditions.
2. *Improved time resolution.* The $N(h)$ conversion represents an amount of arithmetic requiring less time on a fast minicomputer than the echo pulse travel time; thus, the profiles can be produced at virtually any desired rate within this natural limitation.
3. *Acquisition and use of additional information.* Among many examples, of particular relevance is the use of echo angle-of-arrival information to discriminate overhead structure, to improve Auto $N(h)$ echo selection, and to describe tilts and horizontal structure. The Dynasonde accommodates an addressable receiving antenna array and digitizes the complex amplitude of each echo at each antenna to provide such information as part of its Kinesonde functions.

In the larger context of Dynasonde applications, automatic $N(h)$ profiles become considerably more than a desirable data output of the system; they provide essential reference information for adaptive application of the Dynasonde toward most other measurement objectives.

DISCUSSION

Grossi: How do you distinguish between the ordinary and extraordinary returns when they cross?

Wright: If you know the profile you can calculate both curves, and the FINDIT process can be set to the task of finding the information nearest these. This leads then to the identification. If you know the profile, based on previous results, you can predict this. Note that the process can run forward or backward. Also, the angle-of-arrival feature is very important here.

Reinisch: With regard to frequency selection, are you sure of convergence in the loop so you get reasonable frequencies? What do you do if they are unreasonable?

Wright: It is not a fully refined process yet. It is workable, however, as a research tool now.

Grossi: Have you ever computed how much time you have to wait until the cross-correlation coefficient of the ionogram drops to $1/e$?

Wright: You are asking about the autocorrelation time of the electron density at a fixed height and location. No, this has not been done, mainly because of a lack of adequate time resolution and statistics in the past. The Auto $N(h)$ process is a way to gain these statistics, when implemented on a digital Ionosonde.

IONOSPHERIC PROFILE INVERSION USING OBLIQUE-INCIDENCE IONOGRAMS

D. L. Nielson and T. M. Watt

Stanford Research Institute, Menlo Park, California

N73-11607

ABSTRACT

This paper reviews some of the elementary methods used in deriving true-height profiles from oblique-incidence ionograms. The two principal methods presented are oblique-to-vertical transformation and direct inversion of the oblique-incidence ionogram. Limitations in oblique-incidence inversion due to magnetic-field effects, horizontal gradients, and absolute time delay are discussed.

INTRODUCTION

In comparison with the amount of attention given electron-density profile measurement at vertical incidence, use of the equivalent technique at oblique incidence has been almost totally ignored. This avoidance is attributable in part to the lack of facilities with which to perform the measurement, but more to the limitations and inaccuracies encountered. While, in principle, the measurement of any oblique wave can reveal something about the intervening ionosphere, in practice many effects limit the accuracy of what can be inferred from that measurement.

It is useful to define several concepts that are addressed in this paper. There are two sounding geometries — *vertical* and *oblique*. There are two possible operational modes — *monostatic* and *bistatic*. All sounding referred to in this paper takes place in the MF and HF bands of the frequency spectrum (~ 1 MHz to 30 MHz).

Sounding at vertical incidence involves propagating a signal vertically and receiving the reflected signal at the same ground station. Most soundings at these frequencies involve only linearly polarized waves, and a common antenna is used for both transmitting and receiving; this operating mode we refer to as *monostatic*. In certain special cases such as the use of circularly polarized waves, it would be necessary to use separate antennas for transmitting and receiving; this operating mode could be called *bistatic*.

Sounding at oblique incidence involves transmitting a signal such that the wave does not reflect back directly to the transmitter site but rather propagates obliquely through the ionosphere and returns to the earth at a remote location. Monostatic operation at oblique incidence is possible using ground backscatter. In ground backscatter, the signal propagates via the ionosphere and is scattered on reaching the ground. A small fraction of the energy returns to the transmitter/receiver site. A somewhat similar setup in a bistatic mode would result from using a transmitter and receiver at one site and a transponder at a remote site. A transponder located at a remote site receives, amplifies, and retransmits the signal back to the sounding site. The advantages of a transponder are that it provides an amplified return signal and that its location can be precisely determined. The most common oblique sounding mode is bistatic, using a transmitter at one location and a receiver at a remote location.

The reason most often cited for oblique sounding for profile measurement is the greater accessibility to remote areas that it provides. The use of most oblique-incidence sounding, however, has been closely related to operational communication needs, and the area of profile inversion has been of no direct interest to the users.

While ground backscatter can be successfully employed in estimating single parameters such as critical frequency, its use in profile measurement is severely limited. Problems such as measurement of elevation angle, the relatively poor knowledge of the location of ground and ionospheric reflection points, and the extensive ionospheric area involved in typical sweep-frequency operation have prevented its use for this purpose.

The instrumentation for oblique-incidence pulse soundings is not unlike that for vertical-incidence sounding. In the case of monostatic backscatter soundings, the difference could be merely in the antenna; however, other signal-processing measures can be taken to improve the quality of the records. The most satisfactory records for profile measurement at oblique incidence are termed *oblique ionograms*. These records of pulse-group delay are the result of operating a separated sweep- or step-frequency transmitter/receiver pair synchronized in time (and consequently in frequency). Although oblique sounding was originally limited to am pulse waveforms with center frequency swept over the range of propagating frequencies [Agy *et al.*, 1959; Möller, 1964; Wilkens, 1960], synthesized step-frequency pulse [Nielson, 1966], sweep frequency CW (chirp) [Fenwick and Barry, 1966], and extremely wideband high-power pulse, have more recently been used with equipment developed at Stanford Research Institute. Except for transponder systems, all oblique-sounding systems require stable oscillators to maintain timing accuracy. Absolute delay is therefore not normally available.

Ignoring the birefringence of the ionosphere, the lowest order ray at vertical incidence is limited to a single ray at a single frequency. Most ionospheric profiles, on the other hand, give rise at oblique incidence to a high and low ray at frequencies above the vertical-incidence critical frequency. Figure 1 illustrates the relationship between group delays at vertical and oblique incidence. One curve can be considered a transformation of the other. Under certain assumptions, the transformation is very simple and bilateral; without such assumptions, it may be neither. In this paper we discuss these transformations and their attendant assumptions in some detail.

Consider the problem of inverting an oblique ionogram such as that shown in figure 1, under the assumptions of a spherical earth and a concentric, isotropic ionosphere. Using the nomenclature of Croft and Hoogasian [1960], as illustrated in figure 2, we have

$$D = 2r_0 \int_{r_0}^{r_t} \frac{dr}{r \tan \beta}$$

$$= 2 \cos \beta_0 \int_{r_0}^{r_t} \frac{r_0^2 dr}{r(r^2 \mu^2 - r_0^2 \cos^2 \beta_0)^{1/2}} \quad (1)$$

$$P' = 2 \int_{r_0}^{r_t} \frac{dr}{\mu \sin \beta} = 2 \int_{r_0}^{r_t} \frac{r dr}{(r^2 \mu^2 - r_0^2 \cos^2 \beta_0)^{1/2}} \quad (2)$$

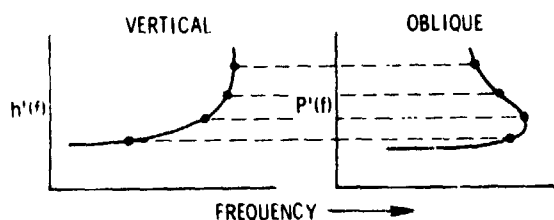


Figure 1.— Correspondence between vertical and oblique-incidence ionograms

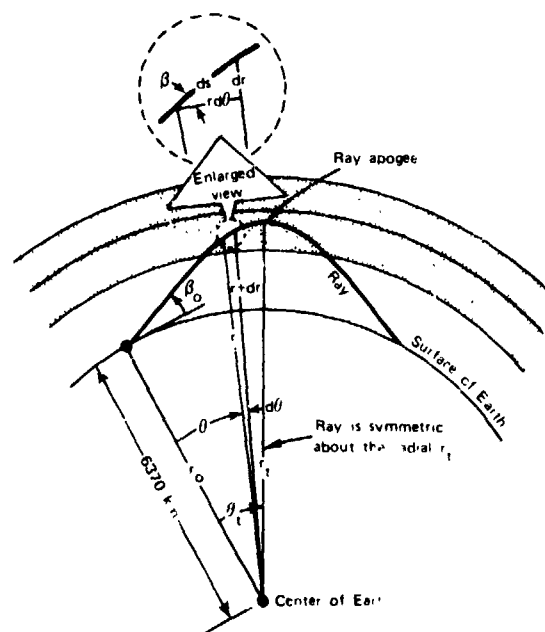


Figure 2.— Nomenclature for an oblique ray path.

$$r\mu \cos \beta = r_0\mu_0 \cos \beta_0 = r_0 \cos \beta_0 \quad (3)$$

where μ is the refractive index at radial distance r . The quantities D and P' are known or measured values, and Eq. (3) is a form of Snell's law (Bouger's rule). For the conditions considered, Eqs. (1) through (3) are exact. From Eq. (3) we obtain the condition satisfied at apogee for any particular frequency:

$$r_t \mu(r_t) = r_0 \cos \beta_0 \quad (4)$$

For an isotropic and collisionless medium, μ is defined by

$$\mu^2(r) = 1 - \frac{f_N^2(r)}{f^2} \quad (5)$$

where f_N is the plasma frequency of the medium, f is sounding frequency, and f_N^2 is proportional to the electron density.

The problem, then, is to determine the dependence of f_N on r that satisfies Eqs. (1) and (2) for all sounding frequencies. We next discuss the capabilities and limitations of the various methods of establishing that relationship.

TRANSFORMATION BETWEEN OBLIQUE AND VERTICAL INCIDENCE

We consider here the direct transformation between corresponding oblique- and vertical-incidence ionograms, without explicitly obtaining an electron-density profile solution — that is, from a $P'(f)$ ionogram, we deduce an $h'_v(f_v)$ ionogram.

Equations (1) and (2) are not, in general, analytically soluble, and tractable procedures can be employed only by making certain approximations. We consider in this section approximations that can be used under appropriate circumstances.

Flat-Earth Flat-Ionosphere Approximation

The flat-earth flat-ionosphere approximation is most easily reached by assuming that the radii r_0 and r_t are much larger than any other dimension. In this case we obtain $r - r_0 = h$, $r_t/r_0 \rightarrow 1$, so that Eqs. (1) through (4) become

$$D = 2 \cos \beta_0 \int_0^{h_t} \frac{dh}{(\mu^2 - \cos^2 \beta_0)^{1/2}} \quad (6)$$

$$P' = 2 \int_0^{h_t} \frac{dh}{(\mu^2 - \cos^2 \beta_0)^{1/2}} \quad (7)$$

$$\mu \cos \beta = \cos \beta_0 \quad (8)$$

$$\mu(h_t) = \cos \beta \quad (9)$$

In the flat-earth flat-ionosphere approximation, a simple equivalence exists between an oblique ionogram and the corresponding vertical ionogram. For any oblique sounding frequency f , we can find the equivalent vertical-sounding frequency, f_v , that corresponds to reflection from the same height, h_f . Comparing Eqs. (5) and (9), we observe that, at the altitude of reflection $f_N = f_{Nr}$, and

$$\cos^2 \beta_O = 1 - \frac{f_{Nr}^2}{f^2} \quad (10)$$

so

$$f_{Nr} = f \sin \beta_O \quad (11)$$

Since, for vertical incidence, $f_{Nr} = f_v$, then

$$f_v = f \sin \beta_O \quad (12)$$

The virtual height corresponding to f_v is found directly from *Marrin's* [1925] theorem and is exactly correct for a flat earth and a flat ionosphere:

$$h'_v = h' = \frac{P'}{2} \sin \beta_O = \frac{D}{2} \tan \beta_O \quad (13)$$

where h'_v is the virtual height associated with a vertical sounding at f_v , and h' is the height of the equivalent triangle associated with an oblique sounding.

Comparison of Eqs. (6) and (7) yields an explicit solution for $\cos \beta_O$ in terms of observed values:

$$\cos \beta_O = \frac{D}{P'} \quad (14)$$

Combining Eqs. (12), (13), and (14), we obtain

$$f_v = f \left[1 - \left(\frac{D}{P'} \right)^2 \right]^{1/2} \quad (15)$$

$$h'_v = \frac{1}{2} \left[(P')^2 - D^2 \right]^{1/2} \quad (16)$$

where, for multiple-hop paths, P' and D are per-hop values. For the flat-earth flat-ionosphere approximation, then, an oblique ionogram can be reduced to a vertical ionogram without introducing any *a priori* assumptions regarding the ionospheric vertical profile. The integral equation

$$h'_v(f_v) = \int_0^{h_f} \frac{dh}{\mu(f_v)} \quad (17)$$

must, of course, still be inverted to solve for the electron-density profile.

Curved-Earth Flat-Ionosphere Approximation

The curved-earth flat-ionosphere approximation is reasonably valid where propagation within the ionosphere does not take place over a large horizontal distance. For this condition, modified forms of the flat-earth flat-ionosphere equations can be used.

For a ground range D as shown in figure 3, the straight-line distance between the end points is D_s , where

$$D_s = 2 r_0 \sin \frac{D}{2r_0} \quad (18)$$

and the height above the chord at the midpoint of D is

$$h'_s = h' + r_0 \left(1 - \cos \frac{D}{2r_0} \right) \quad (19)$$

Defining the angle α_0 to be the angle between the virtual-path ray $P'/2$ and the plane D_s , we can write equations analogous to Eqs. (6), (7), (15), and (16) as

$$D_s = 2 \cos \alpha_0 \int_0^{h_{st}} \frac{dh_s}{(\mu^2 - \cos^2 \alpha_0)^{1/2}} \quad (20)$$

$$P' = 2 \int_0^{h_{st}} \frac{dh_s}{(\mu^2 - \cos^2 \alpha_0)^{1/2}} \quad (21)$$

$$f_v = f \left[1 - \left(\frac{D_s}{P'} \right)^2 \right]^{1/2} \quad (22)$$

$$h'_s = \frac{1}{2} \left[(P')^2 - D_s^2 \right]^{1/2} \quad (23)$$

Thus, using Eqs. (18), (19), (22), and (23), we obtain, for the curved-earth flat ionosphere case,

$$f_v = f \left\{ 1 - \left[\frac{2r_0 \sin (D/2r_0)}{P'} \right]^2 \right\}^{1/2} \quad (24)$$

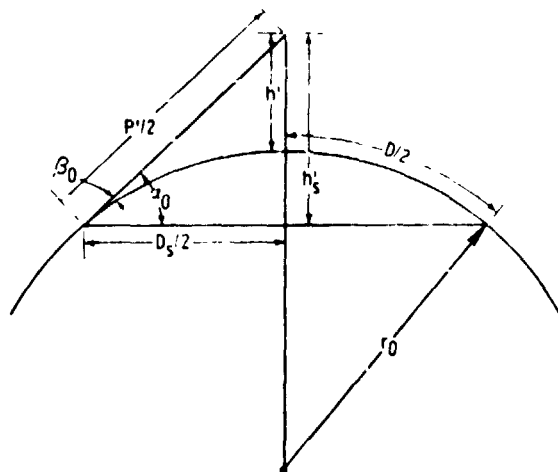


Figure 3.-- Curved-earth geometry.

$$h'_y = h' = \frac{1}{2} \left[(P')^2 - \left(2r_o \sin \frac{D}{2r_o} \right)^2 \right]^{1/2} - r_o \left(1 - \cos \frac{D}{2r_o} \right) \quad (25)$$

where, as before, D and P' are per-hop values. It can be seen that Eq. (24) reduces to Eq. (15), and Eq. (25) reduces to Eq. (16), as P/r_o becomes very small.

Curved-Earth Curved-Ionosphere

Exact relationships. The curved-earth curved-ionosphere case is that described by Eqs. (1) through (4); for this case, there is no tractable, analytical transformation from oblique to vertical incidence without an *a priori* knowledge of the electron-density profile.

In attempting to transform from oblique to vertical incidence, the reasoning proceeds along this line: Assume that for a particular concentric ionosphere of unknown profile an oblique measurement is made of group path P' , at frequency f , over a distance D . Although P' may be multivalued, a single value of P' will correspond to only one ray propagated between the end points. The ray has associated with it some apogee, or maximum real height of propagation into the ionosphere. A group-delay transformation to vertical incidence, then, requires the answers to two questions: At vertical incidence, what sounding frequency would penetrate to the same height as did the obliquely propagated ray? At vertical incidence at such an equivalent sounding frequency, what would be the expected value of virtual height?

A partial answer to the first question can be given immediately. The equivalent vertical-incidence frequency is equal to the ionospheric plasma frequency at the ray apogee. If h_f is the apogee height, then

$$f_v = f_N(h_f) \quad (26)$$

Therefore, from Eq. (5), we have

$$f_v = f \left\{ 1 - \left[\mu(f, f_N) \right]^2 \right\}^{1/2} \quad (27)$$

where the refractive index μ is dependent on the sounding frequency f and the plasma frequency at h_f . By using Eq. (4) and noting that h_f and r_f refer to the same height, we can express f_v in terms of elevation angle:

$$f_v = f \left[1 - \left(\frac{r_o}{r_f} \cos \beta_o \right)^2 \right]^{1/2} \quad (28)$$

Equation (28) is exact, under the stated conditions, but from this point we cannot proceed further without employing approximations.

In the flat-earth flat-ionosphere approximation the problem expressed in Eq. (28) was solved by letting r_o/r_f approach unity. In this way we not only eliminated the effect of the unknown r_f , but we also made the integrands in Eqs. (1) and (2) identical, so that $\cos \beta_o$ could be found as equal to the ratio of the measured quantities D and P' without our having to solve either of the integral equations (1) and (2). In the curved-earth flat-ionosphere approximation, we did essentially the same thing, since the difference between the two approximations is described completely by the trigonometric relationships of Eqs. (18) and (19).

The second question is concerned with determining the virtual height h'_v associated with a vertical sounding at frequency f_v . Under previous assumptions the vertical and oblique virtual heights were easily related by an equivalent triangular path. But, for the curved-earth curved-ionosphere case, it can be seen from Eqs. (1) and (2) that P' and D (or D_s) can be related only through a knowledge of the electron-density profile. Thus, the difficulty has to do with the existence of a right triangle with a base $D_s/2$ and a hypotenuse $P'/2$. Consider figure 3. The takeoff angle α_o could, in principle, be measured for any particular sounding situation, and the three measured quantities would then be interrelated according to

$$\cos \alpha_o = \frac{D_s}{P'} \quad (29)$$

so the height h'_s could be expressed as

$$h'_s = \frac{P'}{2} \sin \alpha_o = \frac{D_s}{2} \tan \alpha_o \quad (30)$$

It can be shown, in fact, that for a curved ionosphere of finite thickness, Eq. (29) is not in general true, since a triangle that relates the measurable quantities cannot be constructed. Consequently, the concepts of virtual height and equivalent virtual path lose their meaning.

Although a profile-independent transformation is not available, some authors [Smith, 1939; Kobayashi, 1971] have attempted to use flat-ionosphere expressions to compensate for the curved ionosphere through the introduction of correction factors. All such correction factors, however, must make some assumptions about the profile itself as well as the angle of incidence onto the ionosphere.

Approximation errors. It has already been stated that, for the general case, an oblique-to-vertical transformation cannot be made without knowledge of the ionospheric layer. In an attempt to assess the seriousness of the flat-ionosphere approximations, a particular model ionosphere has been assumed. The model is a quasiparabolic (QP) layer defined by

$$f_N = \begin{cases} f_{\max} \left[1 - \left(\frac{r - r_m}{y_m} \right)^2 \left(\frac{r_b}{r} \right)^2 \right]^{1/2} & r_b < r < r_m \left(\frac{r_b}{r_b - y_m} \right) \\ 0 & \text{elsewhere} \end{cases} \quad (31)$$

with parameter values $f_{\max} = 10$ MHz, $r_m = 6670$ km, $r_b = 6545$ km, and $y_m = 125$ km. For the particular values chosen, it is intended to approximate a typical daytime F layer. Ionograms were calculated from the layer by using a ray-tracing program [Faul et al., 1968].

The approximation errors associated with the flat-ionosphere approximations (fig. 4) are obtained as follows: By using the model layer and the ray-tracing program, ionograms were generated for vertical incidence and for ground ranges of 1000, 2000, and 3000 km. Neglecting small errors in the numerical ray tracing, the vertical-incidence ionogram was taken to be exactly correct. Next, for each of the three oblique ionograms, equivalent vertical-incidence ionograms were calculated, according to the flat-earth flat-ionosphere [Eqs. (24) and (25)] approximations. Figure 4(a) illustrates the two approximate and the one exact vertical-incidence ionograms for the 1000-km case. Figures 4(b) and (c) illustrates the same comparison for ground ranges of 2000 and 3000 km, respectively.

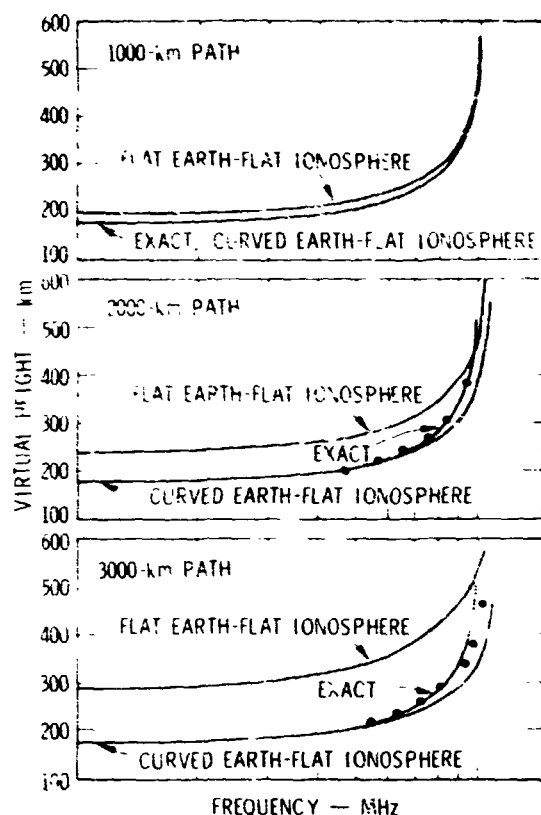


Figure 4. - Typical errors resulting from various oblique-to-vertical transformations.

In an attempt to correct the frequency error introduced by ionospheric curvature, the Newbern Smith correction factor [Davies, 1965] was introduced on the longer paths. The points in figure 4 indicate the result, which in this instance reduced the error to within a few percent.

The comparisons illustrated in figure 4 lead to the following general conclusions:

1. The approximations are best at the shorter ground ranges and become progressively poorer as range increases.
2. Neglect of ionospheric curvature leads to a derived electron-density profile that has a higher critical frequency than the true profile. The error can largely be corrected, but correction is a complicated function of the profile and path length, and a simple factor cannot be expected to work in all instances.
3. At all ranges, the curved-earth flat-ionosphere approximation is much better than the flat-earth flat-ionosphere approximation.
4. The flat-earth flat-ionosphere approximation leads to a derived electron density with altitude higher than that of the actual layer. For this particular model layer, the height error is about 20 km at 1000-km range, 60 km at 2000-km range, and 110 km at 3000-km range.

The usefulness of the approximations depends, of course, on the needs of the user, and the particular results obtained here are only illustrative.

DIRECT INVERSION OF THE OBLIQUE-INCIDENCE EQUATION FOR GROUP DELAY

Lamination Methods

In spite of the additional degree of freedom characteristic of oblique incidence, inversion methods have been developed that are not unlike most of the approaches used at vertical incidence. That is, the ionogram can be inverted by using numerical methods without first converting to vertical incidence. Two of these numerical methods use laminations of electron density across which a particular form is assumed for the variation of electron density, a form that will allow a closed-form solution for the virtual delay and the ground distance [George, 1970; Smith, 1970]. These quantities are then summed across all the laminations, for testing against known or measured values of D and P' . In calculating the plasma frequency for a given lamination (height), the value at the next lowest interval is used. This procedure is common to both the lamination methods, as well as for vertical incidence. Such a procedure relies, of course, on a lack of horizontal gradients, since oblique incidence, over a fixed distance, rays at different frequencies follow different paths.

Of the two laminar methods, the more general is that of *Smith* [1970]. In this method the takeoff angle β_0 and the height of reflection r_t are juggled until P' and the ground distance D are satisfied. To make the search efficient, limits on the takeoff angle are calculated. The variation with electron density across a slab or lamination is given by a hyperbolic relation,

$$f_N^2 = A - \frac{B}{r} \quad (32)$$

which allows a closed form for ground range subtended by the distance the ray is in the slab. The values of A and B are such that f_N^2 is approximately linear in r over small increments in r .

Neglecting horizontal gradients and the magnetic field, the reflection height is given from Eqs. (4) and (5) as

$$r_t = \frac{r_0 \cos \beta_0}{[1 - f_N^2(r)/f^2]^{1/2}} \quad (33)$$

where r_0 is the radius of the earth. This expression is used in conjunction with the appropriate relation

$$P' = \frac{r_t D}{r_0 \cos \beta_0} \quad (34)$$

to obtain, as in Eq. (15),

$$f_N = f \left[1 - \left(\frac{D}{P'} \right)^2 \right]^{1/2} \quad (35)$$

The takeoff angle β_0 must be found by interpolation.

An example of the accuracy available in this type of inversion is given by figure 5 from *Smith* [1970]. Using the electron-density profile on the left (x), he calculated an oblique ionogram by means of ray tracing. Points were then scaled from the oblique trace and an inversion performed. The result is plotted (o) on the vertical profile.

It is apparent that no downward reflection can occur over the height range where a profile is reentrant. While the above "straight-line" approach can be employed as though no reentrance existed, *Smith* shows how to allow for such profiles and what assumptions are required for satisfactory results. Since no portion of the oblique trace can be directly associated with the reentrant region, any information about that region must come from the increase in virtual delay above that height.

George [1970] has suggested a somewhat similar method that follows closely the vertical-incidence work of *Jackson* [1956]. In this instance the angle of the ray path is supplied more explicitly from an empiricized equation

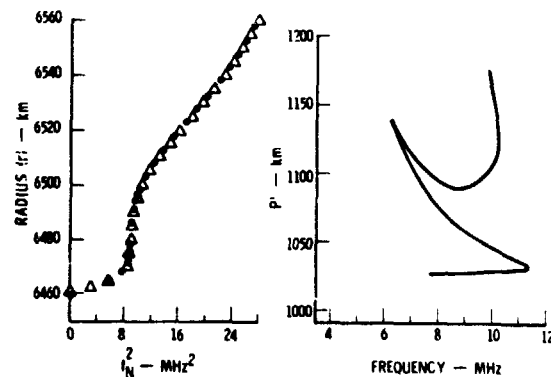


Figure 5. - Example of inversion of oblique ionogram using lamination technique [*Smith*, 1970].

$$\nu_o = \frac{\pi}{2} - \frac{D}{2r_o} - \sin^{-1} \left[\frac{2r_o \sin(D/2r_o)}{P' + \epsilon} \right] \quad (36)$$

rather than obtaining it by interpolation. The term ϵ is a function of the virtual depth into the layer at the frequency of interest. Both of the lamination methods give good results for nonreentrant profiles.

Analytical Models of Layer Shape

Usually we think of profile inversion as the process of determining a more or less completely arbitrary distribution. In many circumstances, however, a more restricted profile is still satisfactory. For these cases, some analytical form for the electron-density profile must be assumed, and the "inversion" consists of evaluating the coefficients or scale factors of the assumed expression. At oblique incidence, two factors make an otherwise straightforward calculation very difficult. First, there is not just a single equation in $P'(f)$, as at vertical incidence, but a set of coupled equations in P' and D , which are easily solved simultaneously only for the case of a flat ionosphere. Second, the accuracy with which the record must be scaled to obtain realistic values of layer parameters often exceeds that available.

Probably the most frequently assumed layer shape for such scaling is parabolic. The three-parameter parabolic model consists of a height, a critical frequency, and a semithickness. Practice has shown that unless data points can be scaled in the vicinity of the junction frequency or on the high ray, the estimates of critical frequency will have substantial error.

A frequently used means for mapping critical frequency over wide geographical areas is the use of ground backscatter. Again, certain assumptions are required for the values or constancy of layer height and thickness. An indication of the type of accuracy and range available by using one-hop ground backscatter is given by *Hatfield* [1970]. The use of ground backscatter data to obtain more complete profile data has been discussed by *Egan* [1960] and *George* [1970]. Measurement difficulties, however, tend to limit the usefulness of the more detailed inversions obtained from backscatter.

LIMITATIONS

Magnetic Field Effects

The inclusion of geomagnetic field effects in propagation calculations makes the equations far more complicated, and for this reason geomagnetic field effects are usually ignored whenever possible. One advantage of oblique (as opposed to vertical) sounding is that the normally higher operating frequencies tend to reduce the effect of the geomagnetic field. When the geomagnetic field is included, it must be recognized that the oblique ray encounters continuous variations in the magnetic field vector along its trajectory. Since it is difficult to treat changing field conditions and yet retain some generality, the angle θ between the ray and the magnetic field is assumed to be constant and equal to, say, an average value in the vicinity of reflection.

The detailed effects of the geomagnetic field on oblique ionograms may be found in *Kopka and Möller* [1968b] for varying path orientations and locations.

Another source of error due to the magnetic field is the lateral displacement of the ray from the no-field trajectory. *Titheridge* [1959] has examined this error and concluded that lateral displacements of the order of 10 km can occur only at frequencies below 5 MHz.

Horizontal Gradients in Electron Density

Perhaps the greatest limitation on the accuracy of inverting oblique-incidence ionograms is the lack of constancy of the electron-density profile over the reflection region. The rays are reflected, of course, in a three-dimensional distribution of electron density, which cannot normally be described by a single vertical profile. Although one might expect to perform the inversion after some correction for known horizontal gradients, such a procedure is difficult to employ (not to mention that the gradients are usually unknown). The inclusion of such gradients would no doubt require tedious trial and error using ray-tracing techniques.

If we cannot treat the gradients quantitatively, we can, however, examine the type of error they might produce in a typical profile. Generally speaking, the maximum oblique frequency tends to be determined near the horizontal region having the maximum electron density. For example, for a simple linear gradient of critical frequency, the maximum-frequency ray will be reflected nearer the end of the path with the higher density. Furthermore, the maximum frequency will exceed the maximum frequency that would be observed if the average critical frequency existed all along the path. Conversely, a simple linear gradient in height will reduce the maximum frequency. *Kopka and Möller [1968a]* examined some specific cases of each type of gradient.

Absolute Time Delay

There are roughly three techniques for measuring absolute path delay by using oblique-incidence sounders. The first, and probably the most accurate, is a simultaneous duplex pulse transmission along the path. The average of the time delays measured at each receiver is the exact path length — correct to the accuracy of the measuring equipment. The use of a transponder falls in this category.

The second method is the use of "constant" time-delay modes, such as those propagated via the normal *E* layer. This method has the obvious disadvantage that the *E*-layer trace is not always detectable, especially over long paths; also, it is subject to *E*-layer height variations, which become more pronounced when the height variation of E_s is included. The accuracy achievable with this method is on the order of 100 μ sec, depending on the length of the path. For a 3100-km path ($2E$), the variation in delay is 100 μ sec as the *E*-layer height is varied from 90 to 130 km.

The third method is the use of separate stable oscillators. This method requires that the standards of the transmitter and those of the receiver be in phase or that their clocks be exactly synchronized. The error in measurement will be the time difference existing between these clocks or the error in the measurement apparatus, whichever is greater. If the actual time difference between clocks is unknown, the path delay will be no more accurate than the computed delay used to establish the original synchronization.

CONCLUSIONS

The inversion of oblique-incidence ionograms can be achieved only under rather restrictive limitations or assumptions. The most serious of these restrictions is the absence of horizontal gradients in electron density. The normal curvature of the ionosphere may be allowed for by using correction factors in the oblique-to-vertical transformation methods or by using the lamination methods. With the assumption of a horizontally stratified ionosphere, many approaches are possible, including oblique-to-vertical transformation or direct inversion of the oblique trace. The inaccuracies due to the magnetic field and to uncertainty in absolute time delay lead to smaller errors for most situations. With care in time-delay scaling and in the selection of oblique-to-vertical transformation methods, a vertical-incidence ionogram can be derived that will accurately portray the electron-density profile if the horizontal gradients are small over the reflection region.

DISCUSSION

Chang: In your flat-earth ionosphere approximation, the profile you got was far from the true one. Could you use the modified refractive index concept to go from that profile to the original profile? *Booker* [1946] used this concept.

Watt: That's similar to the method of Newbern Smith.

Reinisch: The functions you gave for h' seem to all be single valued; for example for one frequency there is one time delay, while the model oblique ionograms show the typical nose. The Peterson ray is attributed to the higher trace. How have you taken care of that?

Watt: The h' functions occur after transforming from oblique to vertical incidence and are single valued.

Unidentified speaker: Is the horizontal gradient problem alleviated if you have several receivers?

Watt: You can learn more about horizontal gradients, but I don't have a quantitative feel for the extent.

Reinisch: Would pulses help to separate the multipath modes compared to a chirp sounder?

Watt: You can separate multipath modes with chirp sounders.

IONOSPHERIC SOUNDING

BIBLIOGRAPHY

- Agy, V., and Davies, K., Ionospheric Investigations Using the Sweep-Frequency Pulse Technique at Oblique Incidence, J. NBS, 63D, 151-174, 1959.
- Agy, V., Davies, K., and Salaman, R., An Atlas of Oblique Incidence Ionograms, National Bureau of Standards Technical Note 31, 1959.
- Alpert, Ya L., Radio Wave Propagation and the Ionosphere, Acad. Sci. Press, Moscow (Russian); Consultants Bureau Enterprises, Inc., New York, 1963.
- Argence, E., Methods of Calculating the True Heights of the Layers of the Ionosphere, Z. Geophys., 20, 91-103, 1954.
- Argence, E., and Mayot, M., A Method for the Determination of the True Heights of Ionospheric Layers, Allowing for the Effect of the Earth's Magnetic Field. I A Method for Using an Approximation for the Refractive Index (Ordinary Ray), J. Geophys. Res., 58, 147-165, June 1953. II. Using the Exact Value of the Refractive Index (Ordinary Ray), J. Geophys. Res., 58, 493-496, Dec. 1953.
- Argence, E., and Mayot, M., Determination of the Height and Semi-Thickness of a Regular Ionospheric Layer Taking into Account the Effect of the Earth's Magnetic Field (Ordinary and Extraordinary Rays), Compt. Rend. (Paris), 238, 1721-1723, April 1954.
- Bauer, S. J., and Jackson, J. E., A Small Multi-Purpose Rocket Payload for Ionospheric Studies, Goddard Space Flight Center document X-615-63-95 presented at the Fourth International Space Sciences Symposium (COSPAR) in Warsaw, Poland, June 3-11, 1963.
- Becker, W., The Analysis of Observations on the Ionosphere, A.E.U., 4, 391-400, Oct. 1950.
- Becker, W., The Calculation of the True Distribution of the Electron Density in the Ionosphere--I, A.E.U., 9, 277-284, March 1955.
- Becker, W., A Method for the Routine Evaluation of the True Distribution of the Electron Density in an Ionospheric Layer From the Ionograms of this Layer, A.E.U., 10, 207-214, Feb. 1956.
- Becker, W., Tabulations of Group Velocities of Ordinary Echoes at Vertical Incidence in the Ionosphere, A.E.U., 11, 166-172, Dec. 1957.
- Becker, W., Tables of Ordinary and Extraordinary Refractive Indices, Group-Refractive Indices and $h'_{o,x}(f)$ Curves for Standard Ionospheric Layer Models, Proc. Max-Planck-Institut für Aeronomie, 4, (Springer, Verlag, Berlin), 1960.
- Becker, W., On the Manual and Digital Computer Methods Used at Lindau for the Conversion of Multifrequency Ionograms to Electron Density-Height Profiles, Radio Sci., 2, 1205-1232, 1967.
- Benson, R. F., Frequency Shifts Observed in the Alouette II Cyclotron Harmonic Plasma Resonances, Proc. IEEE, 57, 1139-1142, 1969.
- Beynon, W. J. G., and Thomas, J. O. The Calculation of the True Heights of Reflection of Radio Waves in the Ionosphere, J. Atmos. Terrest. Phys., 9, 184-200, 1956.

- Bitoun, J., Graff, Ph., and Aubry, M., Ray Tracing in Warm Magnetoplasma and Applications to Topside Resonances, Radio Sci., 5, 1341-1349, 1970.
- Booker, H. G., Application of the Magneto-Ionic Theory to Radio Waves Incident Obliquely Upon a Horizontally-Stratified Ionosphere, J. Geophys. Res., 54, 243-274, 1949.
- Booker, H. G., and Seaton, S. L., Relation Between Actual and Virtual Ionospheric Height, Phys. Rev., 57, 87-94, Jan. 1940.
- Bourdeau, R. E., Aikin, A. C., and Donley, J. L., Lower Ionosphere at Solar Minimum, J. Geophys. Res., 71, 727-740, Feb. 1966.
- Budden, K. G., A Method for Determining the Variation of Electron Density with Height ($N(z)$ curves) from Curves of Equivalent Height Against Frequency ($h'f$ curves), Physics of the Ionosphere, 332-339, Report of Phys. Soc. Conference, London, 1955.
- Budden, K. G., Radio Waves in the Ionosphere, Cambridge Univ. Press, Cambridge, Mass., 1961.
- Cain, J. C., and Cain, S., Derivations of the International Geomagnetic Reference Field (Oct. 1968), NASA Goddard Space Flight Center Rep. X-612-68-501, Dec. 1968.
- Calvert, W., Oblique Z-Mode Echoes in the Topside Ionosphere, J. Geophys. Res., 71, 5579-5584, Dec. 1966.
- Colin, L., Chan, K. L., and Lee, J. G. K., Phase and Group Refractive Indices from the Collisionless Magnetoionic Theory, Radio Sci., 4, 573, June 1969, also NASA TM X-1553, 1968.
- Colin, L., Chan, K. L., and Lee, J. G. K., Computation of Topside Ionograms from $N(h)$ Profiles, Radio Sci., 4, 573, June 1969, also NASA TN D-5052, 1969.
- Cooper, E. A., Phase Integral Calculations of Ionospheric Reflections Including a Variable Collision Frequency, J. Atmos. Terr. Phys., 26, 995-1005, 1964.
- Croft, T. A., Synthesis of Oblique Ionograms by Digital Computer, Stanford Univ. Tech. Rep. 89, Stanford Electronics Lab., Sept. 1964.
- Croft, T. A., and Hoogasian, H., Exact Ray Calculations in a Quasi-Parabolic Ionosphere With No Magnetic Field, Radio Sci., 3, 69-74, 1968.
- Croft, T. A., and Washburn, T. W., Comparison of Simultaneous Forward and Backscatter Soundings to Verify a Proposed Focusing Mechanism of HF Sky Waves, J. Geophys. Res., 74, 2443-2446, May 1969.
- Danilkin, N. P., and Ryss, I. K., Use of Both Components of a Magnetically Split Signal in Analysis of the Ionospheric Effect of the Solar Eclipse of 1961, Geomagn. Aeron., 4, 545-549, 1964.
- Danilkin, N. P., and Zhuchenko, S. A., Computer Calculation of N_z Profiles of the Ionosphere with Allowance for Both Components of a Magnetosplit Signal, Geomagn. Aeron., 4, 239-243, 1964.
- Davies K., Ionospheric Radio Propagation, NBS Mono. 80, Supt. of Documents, Wash. D.C., April 1, 1965.

- Davies, K., and Sahai, A. K., Study of the 'Valley Problem' with a Ray Tracing Program, Electron Density Profiles in the Ionosphere and Exosphere, B. Maehlum, ed., Pergamon Press, N.Y., 162-166, 1962.
- Doupnik, J. R., A Flexible Method of Determining the Electron Density in the Ionosphere, Sci. Rep. 190, Ionosphere Res. Lab., Penn. State Univ., University Park, Pa., 1963.
- Doupnik, J. R., and Schmerling, E. R., The Reduction of Ionograms from the Bottomside and Topside, J. Atmos. Terr. Phys., 27, 917-941, 1965.
- Duncan, R. A., Computations of Electron Density Distributions in the Ionosphere Making Full Allowance for the Geomagnetic Field, J. Geophys. Res., 63, 491-499, 1958.
- Egan, R. D., Some Ionospheric Results, in IGY, W.J.G. Beynon, ed., Elsevier, N.Y., 231, 1960.
- Farmer, F. T., Childs, C. B., and Cowie, A., Critical Frequency Measurements of Wireless Waves Reflected Obliquely from the Ionosphere, Proc. Phys. Soc. (London), 50, 767, 1938.
- Franklin, C. A., and Maclean, M. A., The Design of Swept Frequency Topside Sounders, Proc. IEEE, 57, 897-929, 1969.
- George, P. L., The Application of Model Technique to the Reduction of Oblique Ionospheric Data, Tech. Note CPD(T), 147, Dept. of Supply, Weapons Res. Establish., Salisbury, South Australia, Dec. 1967.
- George, P. L., True Height Analysis of Oblique Incidence H.F. Radio Wave Data, J. Atmos. Terr. Phys., 32, 905-916, 1970.
- Gething, P. J. D., Lateral Deviation and Effective Path Length for Oblique Propagation Through the Ionosphere, Proc. Int'l Conf. Ionosphere, Inst. of Phys. and Phys. Soc., London, July 1962.
- Gething, P. J. D., and Maliphant, R. G., Unz's Application of Schlomilch's Integral Equation to Oblique Incidence Observations, J. Atmos. Terr. Phys., 29, 599-600, 1967.
- Goubau, G., Connection Between the Apparent and the True Height of the Ionosphere Allowing for Magnetic Bi-Refringence, Hochfreq. Elek., 44, 17-23, 1934.
- Goubau, G., On the Dispersion Theory of the Ionosphere, Hochfreq. Elek., 45, 179-185, 1935.
- Grawert, G., and Lassen, H., The Interpretation of Ionospheric Observations, Z. Angew. Phys., 6, 136-143, 1954.
- Guha, U. C., Analysis of Virtual Height-Frequency Records, J. Geophys. Res., 54, 355-362, Dec. 1949.
- Hagg, E. L., Electron Densities of 8-100 Electron cm^{-3} Deduced from Alouette II High Latitude Ionograms, Can. J. Phys., 45, 27-36, 1967.
- Hagg, E. L., Hewens, E. J., and Nelms, G. L., The Interpretation of Topside Sounder Ionograms, Proc. IEEE, 57, 949-960, 1969.
- Hatfield, V. E., Derivation of Ionospheric Parameters From Backscatter Data, in Ionospheric Forecasting AGARD Conf. Proc., No. 49, 1970.

- Herbert, Th., Tables of Virtual Heights for Models of Monotonic and Nonmonotonic Ionospheric Layers, Radio Sci., 2, 1269-1277, 1967.
- Hojo, H., The Direct Use of the Phase Refractive Index for Reducing $h'(f)$ curves to $N(h)$ Profiles, Radio Sci., 2, 1177-1193, 1967.
- Howe, H. H., and McKinnis, D. E., Ionospheric Electron-Density Profiles with Continuous Gradients and Underlying Ionization Corrections: II. Formulation for a Digital Computer, Radio Sci., 2, 1135-1158, 1967.
- Jackson, J. E., A New Method for Obtaining Electron Density Profiles from $p'f$ Records, J. Geophys. Res., 61, 107-127, 1956.
- Jackson, J. E., The Analysis of Topside Ionograms, Goddard Space Flight Center Rept. X-615-67-452, 1967.
- Jackson, J. E., The Reduction of Topside Ionograms to Electron-Density Profiles, Proc. IEEE, 57, 960-976, June 1969. a.
- Jackson, J. E., Comparisons Between Topside and Ground-Based Soundings, Proc. IEEE, 57, 976-985, June 1969. b.
- Kelso, J. M., A Procedure for the Determination of the Vertical Distribution of the Electron Density in the Ionosphere, J. Geophys. Res., 57, 357-367, Sept. 1952. Also Penn. State Univ. Ionospheric Res. Sci. Rep. No. 19, 1952.
- Kelso, J. M., The Determination of the Electron Density Distribution of an Ionosphere Layer in the Presence of an External Magnetic Field, J. Atmos. Terr. Phys., 5, 11-27, 1954.
- Kelso, J. M., The Calculation of Ionospheric Electron Density Distributions, J. Atmos. Terr. Phys., 10, 103-109, 1957.
- Kelso, J. M., Radio Ray Propagation in the Ionosphere, McGraw-Hill Book Co. Inc., N.Y., 1964.
- King, G. A. M., Electron Distribution in the Ionosphere, J. Atmos. Terr. Phys., 5, 245-246, 1954.
- King, G. A. M., Electron Distribution in the Ionosphere, J. Atmos. Terr. Phys., 8, 184-185, 1956.
- King, G. A. M., Relation Between Virtual and Actual Heights in the Ionosphere, J. Atmos. Terr. Phys., 11, 209-222, 1957.
- King, G. A. M., Use of Logarithmic Frequency Spacing in Ionogram Analysis, J. Res. NBS 64D (Radio Prop.), 5, 501-504, 1960.
- Kobayashi, T., Transmission Curves for the Curved Ionosphere, J. Radio Res. Lab. (Japan), 8, 395-411, 1961.
- Kopka, H., and Moller, H. G., Interpretation of Anomalous Oblique Incidence Sweep-Frequency Records Using Ray Tracking, Radio Sci., 3, 43-52, 1968. a.
- Kopka, H., and Möller, H. G., MUF Calculations Including the Effect of the Earth's Magnetic Field, Radio Sci., 3, 53-56, 1968. b.
- Kraus, L., Further Discussion of Kelso's Paper: On a Method for Determination of the Distribution of Electron Density in the Ionosphere, J. Geophys. Res., 58, 551-553, Dec. 1953.

- Kutiyeu, I. S., A Method of Determining the Electron Density Distribution in the Ionosphere, Geomagn. Aeron., 10, 433, 1970.
- Lockwood, G. E. K., A Computer-Aided System for Scaling Topside Ionograms, Proc. IEEE, 57, 986-989, 1969.
- Lockwood, G. E. K., A Modified Iteration Technique for Use in Computing Electron Density Profiles from Topside Ionograms, Radio Sci., 5, 575-578, 1970.
- Long, A. R., and Thomas, J. O., Titheridge Coefficients for the Polynomial Method of Deducing Electron Density Profiles from Ionograms, J. Res. NBS 67D (Radio Prop.), 1, 7-82, 1963.
- Madsen, R. G., Closed-Loop Processing of Topside Ionospheric Sounder Data--Study Report Phase II, NASA CR 73429, 1970.
- Mal'tseva, O. A., Method of Correcting the N(h) Profiles of the F-Region with Allowance for a 'Valley', Geomagn. Aeron., IX, 753-755, 1969.
- Manning, L. A., The Determination of Ionospheric Electron Distribution, Proc. IRE, 35, 1203-1207, Nov 1947.
- Manning, L. A., The Reliability of Ionospheric Height Determinations, Proc. IRE, 37, 599-603, 1949.
- Manning, L. A., Discussion of Kelso's Paper: On a Procedure for the Determination of the Vertical Distribution of the Electron Density in the Ionosphere, J. Geophys. Res., 58, 117-118, March 1953.
- Mariani, F., Electron Density in the Upper Atmosphere and the Interpretation of $h'(f) \cos$, Ann. Geofis., 6, 21-45, Jan. 1953.
- Martyn, D. F., The Propagation of Medium Radio Waves in the Ionosphere, Proc. Phys. Soc., 47, 323-339, 1935.
- Millington, G., Attenuation and Group Retardation in the Ionosphere, Proc. Phys. Soc., 50, 561-580, July 1938.
- Millington, G., The Relation Between Ionospheric Transmission Phenomena at Oblique and Those at Vertical Incidence, Proc Phys. Soc., 50, 801-825, 1938.
- Misyura, V. A., Krokhmal'nikov, Ye. B., Zinchenko, G. N., Biryukov, G. A., and Gridin, A. N., Contribution to the Determination of Ionospheric Radio Wave Propagation Effects from the Faraday and Doppler Effects of Coherent Signals from Geophysical Rockets Recorded at Spaced Points, Geomagn. Aeron., 10, 335-340, 1970.
- Möller, H. G., Variable-Frequency Pulse Transmission Tests at Oblique Incidence Over Distances Between 1000 and 2000 km, Authorized Translation by Stanford Research Institute, Menlo Park, Calif., publisher, Westdeutscher Verlag, Cologne, July 1964.
- Muldrew, D. B., An Ionospheric Ray-Tracing Technique and Its Application to a Problem in Long-Distance Radio Propagation, IRE Trans. Ant. Prop., AP-7, 393-396, Oct. 1959.
- Muldrew, D. B., and Maliphant, R. G., Long Distance One-Hop Ionospheric Radio-Wave Propagation, J. Geophys. Res., 65, 1805-1815, 1962.

- Muldrew, D. B., Radio Propagation Along Magnetic Field-Aligned Sheets of Ionization Observed by the Alouette Topside Sounder, J. Geophys. Res., **68**, 5355-5370, 1963.
- Murray, F. H., Erratum: Heights of Reflection of Radio Waves in the Ionosphere, Phys. Rev., **51**, 779, May 1937.
- Murray, F. H., and Hoag, J. B., Heights of Reflection of Radio Waves in the Ionosphere, Phys. Rev., **51**, 333-341, March 1937.
- Nielson, D. L., and Hagn, G. H., Frequency Transformation Techniques Applied to Oblique-Incidence Ionograms, Res. Mem. 13, DASA Contract DA-36-039-SC-87197, Stanford Research Institute, Menlo Park, Calif., January 1964.
- Nielson, D. L., Spread-F and Its Effects Upon Radio Wave Propagation and Communication, AGARDograph 95, Technivision, Slough, England, 1966.
- Ostrowski, A. M., Solution of Equations and Systems of Equations, Academic Press, 1960.
- Paul, A. K., Bestimmung der wahren aus der scheinbaren Reflexionshöhe, AEU, **14**, 468, 1960.
- Paul, A. K., Ionospheric Electron-Density Profiles with Continuous Gradients and Underlying Ionization Corrections. I. The Mathematical Physical Problem of Real-Height Determination from Ionograms, Radio Sci., **2**, 1127-1133, 1967.
- Paul, A. K., The Use of Virtual-Height Slopes for the Determination of Electron-Density Profiles, Radio Sci., **2**, 1195-1204, 1967.
- Paul, A. K., and Smith, G. H., Generalization of Abel's Solution for Both Magnetoionic Components in the Real-Height Problem, Radio Sci., **3**, 163-170, 1968.
- Paul, A. K., Smith, G., and Wright, J. W., Ray-Tracing Synthesis of Ionogram Observations of a Large Local Disturbance in the Ionosphere, Radio Sci., **3**, 15-26, 1968.
- Paul, A. K., and Wright, J. W., Some Results of a New Method for Obtaining Ionospheric N(h) Profiles and Their Bearing on the Structure of the Lower F Region, J. Geophys. Res., **68**, 5413-5420, 1963.
- Paul, A. K., and Wright, J. W., Electron Density Profile Analysis of Topside Sounder Ionograms, J. Geophys. Res., **69**, 1431-1435, 1964.
- Pierce, J. A., The True Height of an Ionospheric Layer, Phys. Rev., **71**, 698-700, May 1947.
- Piggott, W. R., and Rawer, K., ed., URSI Handbook of Ionogram Interpretation and Reduction, Elsevier Publ. Co., Amsterdam, 1961.
- Plamondon, J., Sifford, B., and Temple, M. V., HF Communication Effects: Automatic Analysis of Oblique-Incidence Sounder Data, DASA 1705, DASA-Contract 36-039 SC-87197, Stanford Research Institute, Menlo Park, Calif., August 1965.
- Rao, N. N., Swarm, H. M., and Harrison, A. E., Experimental Determination of the Absolute Phase Path of Radio Waves Reflected from the Ionosphere at Normal Incidence, Proc. IEEE, **52**, 858-859, 1964.
- Ratcliffe, J. A., A Quick Method for Analysing Ionospheric Records, J. Geophys. Res., **56**, 463-485, Dec. 1951.
- Ratcliffe, J. A., The Magneto-Ionic Theory and Its Applications to the Ionosphere, Cambridge Univ. Press, Cambridge, 1959.

- Rawer, K., and Taieb, C., Determination Des Profils De Densité Electronique Dans l'Ionosphere (cas ou l'on tient compte du champ magnétique terrestre) Compt. Rend., 251 760-761, Aug. 1960.
- Rydbeck, O. E. H., The Propagation of Electromagnetic Waves in an Ionized Medium and the Calculation of the True Height of the Ionized Layers of the Atmosphere, Phil. Mag., 30, 282-293, Oct. 1940.
- Rydbeck, O. E. H., Further Notes on the Electron Density Distribution of the Upper Atmosphere, Terr. Mag. Atmos. Elec., 47, 219-226, Sept. 1942.
- Rydbeck, O. E. H., A Theoretical Survey of the Possibilities of Determining the Distribution of the Free Electrons in the Upper Atmosphere, Trans. Chalmers Tech., No. 3, 1942.
- Rydbeck, O. E. H., The Reflection of Electromagnetic Waves From a Parabolic Friction Free Ionized Layer, J. Appl. Phys., 13, 577-581, Sept. 1942.
- Saha, A. K., On the Determination of Electron Density Distribution in the Ionospheric Regions from $h'(f)$ Curves, Ind. J. Phys., 30, 464-479, Sept. 1956.
- Saha, A. K., Validity of Ionospheric True Height Analyses Using the Combined Ordinary and Extraordinary Ray Traces, Ind. J. Pure Appl. Phys., 2, 242-249, 1964.
- Schmerling, E. R., The Reduction of $h'(f)$ Records to Electron-Density-Height Profiles, Penn. State Univ., Ionospheric Res. Sci. Rep. No. 94, June 1957.
- Schmerling, E. R., An Easily Applied Method for the Reduction of $h'(f)$ Records to $N(h)$ Profiles Including the effects of the Earth's Magnetic Field, J. Atmos. Terr. Phys., 12, 8-16, 1958. (Also published as Sci. Rep. 96, Penn. State Univ. Ionosphere Res. Lab., 1957).
- Schmerling, E. R., The Reduction of Ionograms to Electron Density Profiles, J. Atmos. Terr. Phys., 25, 609, 1963.
- Schmerling, E. R., Ten-Point Method of Ionogram Reduction, Radio Sci., 2, 1233-1236, 1967.
- Schmerling, E. R., and Ventrice, C. A., Coefficients for the Rapid Reduction of $h'-f$ Records to $N-h$ Profiles without Computing Aids, J. Atmos. Terr. Phys., 14, 249-261, 1959.
- Seddon, J. C., and Jackson, J. E., Ionosphere Electron Densities and Differential Absorption, Ann. Geophys., 14, 456-463, Oct.-Dec. 1958.
- Shapiro, B. S., and Vinnikova, T. L., The Programming of Computations of Ionospheric True Heights on Electronic Computers, Geomagn. Aeron., 3, 218-223, 1963.
- Shinn, D. H., The Analysis of Ionospheric Records (Ordinary Ray)-I, J. Atmos. Terr. Phys., 4, 240-254, 1953.
- Shinn, D. H., Further Remarks on Kelso's Paper: A Procedure for the Determination of the Vertical Distribution of the Electron Density in the Ionosphere, J. Geophys. Res., 58, 416-418, Sept. 1953.

Shinn, D. H., Tables of Group Refractive Index in the Ionosphere for the Ordinary Ray, Rep. Cambridge Conf. Ionospheric Phys. (Phys. Soc. London), 404-406, 1954.

Shinn, D. H., and Whale, H. A., Group Velocities and Group Heights from the Magneto-Ionic Theory, J. Atmos. Terr. Phys., 2, 85-105, 1952.

Shmoys, J., and Karp, S. N., Calculation of Charge Density Distribution of Multilayers from Transit Time Data, J. Geophys. Res., 61, 183-191, June 1956.

Sifford, B. M., HF Communication Effect Simulation: Computer Processing of Oblique Sounder Data, DASA 1492, Defense Atomic Support Agency Contract DA-36-039 SC-89197, Stanford Research Institute, Menlo Park, Calif., Feb. 1964.

Silberstein, R., The Use of Sweep-Frequency Backscatter Data for Determining Oblique Incidence Ionospheric Characteristics, J. Geophys. Res., 63, 335-351, 1958.

Smith, M. S., The Calculation of Ionospheric Profiles from Data Given on Oblique Incidence Ionograms, J. Atmos. Terr. Phys., 32, 1047-1056, 1970.

Smith, N., The Relation of Radio Skywave Transmission to Ionosphere Measurements, Proc. IRE, 27, 332-347, 1939.

Storey, L. R. O., The Joint Use of the Ordinary and Extraordinary Virtual Height Curves in Determining Ionospheric Layer Profiles, J. Res. NBS 64D (Radio Prop.), 2, 111-124, 1960.

Taieb, C., Method of Determination of Profiles of Electronic Density in the Ionosphere by the Means of Parameters Read on the Ionograms, Ann. Geophys., 18, 227-251, 1962.

Taieb, C., A Quick Model Method for Obtaining Real-Height Parameters from Routine Ionospheric Data, Radio Sci., 2, 1263-1267, 1967.

Thomas, J. O., The Distribution of Electrons in the Ionosphere, Proc. IRE, 47, 162-175, 1959.

Thomas, J. O., Long, A. R., and Westover, D., The Calculation of Electron Density Profiles From Topside Sounder Records, J. Geophys. Res., 68, 3237-3242, 1963.

Thomas, J. O., and Vickers, M. D., The Conversion of Ionospheric Virtual Height-Frequency Curves to Electron Density-Height Profiles, DSIR Radio Res. Spec. Rep. No. 28, 1958.

Titheridge, J. E., Ray Paths in the Ionosphere—Approximate Calculations in the Presence of the Earth's Magnetic Field, J. Atmos. Terr. Phys., 14, 50-62, 1959.

Titheridge, J. E., The Calculation of Real and Virtual Heights in the Ionosphere, J. Atmos. Terr. Phys., 17, 88-109, 1959.

Titheridge, J. E., The Use of the Extraordinary Ray in the Analysis of Ionospheric Records, J. Atmos. Terr. Phys., 17, 110-125, 1959.

Titheridge, J. E., A New Method for the Analysis of Ionospheric $h'(f)$ Records, J. Atmos. Terr. Phys., 21, 1-12, 1961.

Titheridge, J. E., The Analysis of Nighttime $h'(f)$ Records, J. Atmos. Terr. Phys., 20, 209-210, 1961.

- Titheridge, J. E., The Analysis of Ionospheric $h'(f)$ Records Using the Phase Refractive Index, J. Atmos. Terr. Phys., 25, 43-47, 1963.
- Titheridge, J. E., The Calculation of the Heights of the Peaks on the Ionospheric Layers, J. Atmos. Terr. Phys., 28, 267-269, 1966.
- Titheridge, J. E., Calculation of the Virtual Height and Absorption of Radio Waves in the Ionosphere, Radio Sci., 2, 133-138, 1967.
- Titheridge, J. E., Direct Manual Calculations of Ionospheric Parameters Using a Single-Polynomial Analysis, Radio Sci., 2, 1237-1253, 1967.
- Titheridge, J. E., The Overlapping-Polynomial Analysis of Ionograms, Radio Sci., 2, 1169-1175, 1967.
- Tushentsova, I. A., Influences of Sphericity on the Determination of the Horizontal Gradients of Ionospheric Parameters from Oblique Sounding Data by the Method of Equivalent Slopes, Geomagn. Aeron., 10, 57, 1970.
- Unz, H., A Solution of the Integral Equation $h'(f) = \int \mu'(f; f_0) dz(f_0)$, J. Atmos. Terr. Phys., 21, 40-45, 1961.
- Unz, H., Schlomilch's Integral Equation for Oblique Incidence, J. Atmos. Terr. Phys., 28, 315-316, 1966.
- Whale, H. A., and Stanley, J. P., Group and Phase Velocities from the Magneto-Ionic Theory, J. Atmos. Terr. Phys., 1, 82-94, 1950.
- Whale, H. A., Determination of Electron Densities in the Ionosphere from Experimental $h'(f)$ Curves, J. Atmos. Terr. Phys., 1, 244-253, 1951.
- Wilkins, A. F., HF Propagation: Its Present and Future for Communication Purposes, J. Br. Inst. Radio Engrs., 20, 939, 1960.
- Wright, J. W., Ionospheric Electron-Density Profiles with Continuous Gradients and Underlying Ionization Corrections: III. Practical Procedures and Some Instructive Examples, Radio Sci., 2, 1159-1168, 1967.
- Wright, J. W., Some Current developments in Radio Systems for Sounding Ionospheric Structure and Motions, Proc. IEEE, 57, 481-486, 1969.
- Wright, J. W., Knecht, R. W., and Davies, K., Methods of True-Height Analysis, Ann. IGY, 3, 62-68, 1957.
- Wright, J. W., and Smith, G. H., A Review of Current Methods for Obtaining Electron Density Profiles from Ionograms, Radio Sci., 2, 1119-1125, 1967.

5. PARTICLE SCATTERING: CLASSICAL AND QUANTUM MECHANICAL

This chapter is concerned with the scattering of particles, either by each other or by some scattering center. In contrast to the previous chapters, the probing "signals" are particle beams, at least for the classical case, and the measured data are the number of scattered particles emerging per unit time in a given direction, per unit solid angle. The aim of the inverse problem is to infer the interparticle force, or potential energy, as a function of distance from the scattering center.

In general, the problem here is more complex and difficult than those in the other disciplines. The mathematical details and inversion techniques are significantly more sophisticated. Curiously, the laboratory experimental phase of this work is not, to date, as sophisticated as the theoretical work, but the mathematical methods described here have been applied to several other inversion disciplines, in particular to electromagnetic scattering (Chapter 6) and to seismology (Chapter 7).

R. G. Newton organized and chaired the session devoted to particle scattering.

INTRODUCTION TO THE INVERSE QUANTUM SCATTERING PROBLEM*

Roger G. Newton

Indiana University

N73-11608

ABSTRACT

The inverse scattering problem is introduced in the context of the Schrödinger equation. The emphasis is on the Gel'fand-Levitan approach to the nonrelativistic problem at fixed angular momentum.

INTRODUCTION

The experimental situation under discussion in this chapter is, schematically, the following. A colimated beam of particles of given momentum is directed at a thin foil of target-particles that may be considered at rest. The force exerted by the target on the beam particles deflects the latter from their original direction, and they emerge at various angles where they are counted by detectors. The ratio of the number of scattered particles emerging per unit time in a given direction, per unit solid angle, to the incident flux and to the target density, is called the *differential scattering cross section*. It is theoretically most directly described in the center-of-mass coordinate system of the missile and target particles, rather than in the laboratory system in which initially the target is at rest. The two are connected by a simple transformation.

The dynamics of the interaction between the beam and target particles may be described by classical mechanics if the beam energy is high enough and not sufficiently well defined, or more generally by quantum mechanics. In either case the aim of the inverse problem is to infer from the scattering data (the differential cross section) the interparticle force, or potential energy, as a function of distance (the "profile"). The "profile" in this case is not generally expected to exhibit as much detail and small-scale structure as in the cases discussed in previous chapters. On the other hand, the path from the data to the function sought is very much longer and the inference more indirect. As a consequence, the mathematical apparatus needed is much more sophisticated.

The inverse scattering problem in classical mechanics was treated in a paper by *Keller et al.* [1956] and it leads to Abel's integral equation, which we have already seen a number of times. In view of the formal similarity between classical mechanics and ray optics, the appearance of Abel's equation in both is no surprise. Here we turn directly to the quantum mechanical dynamics, and to simplify matters, we shall restrict the discussion to particles without spin at nonrelativistic velocities.

QUANTUM - MECHANICAL DYNAMICS

The scattering of a beam of spinless particles by a rotationally symmetric center of force is described by the time-independent Schrödinger equation, which in appropriate units reads

*Work supported in part by the National Science Foundation and the U.S. Army Research Office, Durham, North Carolina.

$$[-\nabla^2 + V(r)] \psi = E\psi$$

where ∇^2 is the Laplace operator in three dimensions, E is the energy of the particle, and $V(r)$ is the potential energy, assumed spherically symmetric and rapidly decreasing to zero at large r . The asymptotic form of the solution ψ for large r is

$$\psi \sim e^{i\mathbf{k} \cdot \mathbf{r}} + r^{-1} e^{ikr} f(\mathbf{k}', \mathbf{k})$$

where $\hbar\mathbf{k}$ is the momentum of the particle, and $\mathbf{k}' = \mathbf{k}\mathbf{r}/r$. (We write r for $|\mathbf{r}|$.) The connection between the energy and the value of $k = |\mathbf{k}|$ is (in our units) $E = k^2$. You may, of course, equally well think of E as the frequency of the wave, and of \mathbf{k} as the wave vector, with the wave length $\lambda = 2\pi/k$.

The measured differential cross section for scattering from a beam in the direction \mathbf{k} to a counter in the direction \mathbf{k}' is given by

$$\frac{d\sigma}{d\Omega} = |f(\mathbf{k}', \mathbf{k})|^2$$

Owing to the rotational invariance of the scattering center, f depends only on E and on the angle θ between \mathbf{k} and \mathbf{k}' :

$$f(\mathbf{k}', \mathbf{k}) = A(E, x) \quad x = \cos \theta = \frac{\mathbf{k}' \cdot \mathbf{k}}{E}$$

The usual procedure is to solve the Schrödinger equation in some way—exactly, numerically, or approximately—to extract f from ψ , and thereby to calculate the cross section. The inverse scattering problem is posed by assuming $d\sigma/d\Omega$ to be given as a function of E , or x , or both, and then to ask whether there is a (reasonable) function $V(r)$ that produces it via the Schrödinger equation, and if so, whether it is unique and how it can be calculated.

Assuming $d\sigma/d\Omega$ to be given as a function of the angle or x for $0 \leq x \leq 1$, the first problem is to determine the complex function $A(E, x)$ from a knowledge of its absolute magnitude. The most direct attack on this problem is made by means of the so-called “generalized optical theorem,” or unitarity of the S matrix. It follows from the conservation of flux implied by the Schrödinger equation that

$$\text{Im}f(\mathbf{k}', \mathbf{k}) = \frac{k}{4\pi} \int d\Omega'' f^*(\mathbf{k}'', \mathbf{k}') f(\mathbf{k}'', \mathbf{k})$$

where $d\Omega''$ is the solid-angle element of \mathbf{k}'' , and Im denotes the imaginary part of the complex function. If we write (at fixed E)

$$kA(E, x) = F(x)e^{i\phi(x)} \quad 0 \leq F(x)$$

then this equation reads

$$F(x) \sin \phi(x) = \frac{1}{2\pi} \int dy dz \frac{F(y)F(z) \cos [\phi(y) - \phi(z)]}{(1 - x^2 - y^2 - z^2 + 2xyz)^{1/2}}$$

and the integration extends over the interior of the ellipse where $1 - x^2 - y^2 - z^2 + 2xyz \geq 0$. This equation may be regarded as a nonlinear integral equation for $\sin \phi(x)$ if $F(x)$ is given. [Martin, 1969; Newton, 1968]. Both Schauder's fixed-point theorem and the principle of contraction mapping have been applied to this problem, with the following result. If

$$Q(x) \equiv \int dy dz \frac{F(y)F(z)}{2\pi F(x)(1 - x^2 - y^2 - z^2 + 2xyz)^{1/2}} < 1$$

for all x , $0 \leq x \leq 1$, and $F(x)$ satisfies a Hölder condition, then a solution $\sin \phi(x)$ exists. If $Q(x) < 2^{1/2}$, the solution is unique and may be obtained by iteration. [It was shown by Martin, 1969, that for $Q(x) < 0.79$ the solution is unique.] It is also known from explicit examples [Crichton, 1966] that for some $F(x)$ with $Q(x) > 1$ a solution exists and is not unique. The general situation for $Q(x)$ is unknown. It is also not known whether the gap between $2^{1/2}$ (or 0.79) and 1 is real or merely a technical difficulty. For further remarks, see Eftimiu [1970], Goldberg [1970], and Gerber and Karplus [1970, 1971].

Let us assume now that the scattering amplitude $A(E, x)$ has been obtained from the experimental data. We may expand its dependence on x in a Legendre series:

$$A(E, x) = \sum_{\ell=0}^{\infty} a_{\ell}(E) P_{\ell}(x)$$

in which the $P_{\ell}(x)$ are Legendre polynomials. The principal virtue of this expansion is that it allows us to incorporate the restriction imposed on $A(E, x)$ by the generalized optical theorem in a very simple manner: The coefficients a_{ℓ} must have the form

$$a_{\ell} = (2\ell + 1)(ik)^{-1} (e^{2i\delta_{\ell}})$$

where δ_{ℓ} is a real number. It is called the ℓ th phase shift and, of course, depends on E . To say that $A(E, x)$ is given is therefore equivalent to saying that the phase shifts are given.

At this point the inverse problem may be posed in two different ways. The first is to assume that a single phase shift $\delta_{\ell}(E)$ is given as a function of E , for all positive E . The second is to suppose that at one value of $E \neq 0$, all phase shifts are given. The first was attacked much earlier and there is much more literature on it (see appendix) than on the second, even though from a practical point of view the second way of posing the inverse scattering problem is of greater interest. I shall discuss only the first, the so-called "inverse scattering problem at fixed angular momentum." The second, the "inverse scattering problem at fixed energy," is discussed by Sabatier in the next paper.

Before proceeding, there is a more general remark that should be made. It would be most natural to ask: What if $A(E, x)$ were given as a function of both E and x . Or, what if all phase shifts were given as functions of E for all positive E ? We shall see that, at least within a certain rather large class of potentials, a single phase shift as a function of E determines an n -parameter family of potentials (with $n < \infty$). Hence the prescription of $\delta_{\ell}(E)$

for one ℓ determines, for each $\ell' \neq \ell$, at most an n -parameter family of phase-shift functions $\delta_{\ell'}(E)$. Thus while $\delta_{\ell}(E)$, for any one ℓ can be prescribed (within wide limits) arbitrarily as a function of E , two phase shifts of different ℓ values are very much restricted in their possible E dependence by the requirement that they be produced by the same potential.

A similar situation exists for the inverse problem at fixed energy. As Sabatier notes, a set of phase shifts at one energy determines a certain class of potentials. A set of phase shifts at another energy also determines a class of potentials. The intersection of these two classes, for arbitrarily prescribed sets of phase shifts is not known, but the chances are that with probability 1 this intersection will be zero. In other words, when all $\delta_{\ell}(E)$ are given for all E , a horrendous existence problem arises. For this reason, one usually assumes either all δ_{ℓ} at one E , or one δ_{ℓ} for all E .

THE INVERSE PROBLEM AT FIXED ANGULAR MOMENTUM

We are concerned here with one particular general approach to the inverse problem at fixed angular momentum. There is another class of methods more suitable to the extraction of partial information from partial input, rather than from the full knowledge of a $\delta_{\ell}(E)$ for all positive E as discussed by Calogero in a later paper.

If the solution $\psi(r)$ of the Schrödinger equation is subjected to Legendre expansion of its directional dependence on r relative to the direction \hat{k} of the incoming beam

$$\psi(r) = \sum_{\ell=0}^{\infty} (2\ell+1) r^{-1} \psi_{\ell}(r) P_{\ell}(k \cdot r)$$

then the radial function $\psi_{\ell}(r)$ must satisfy the radial Schrödinger equation

$$-\psi_{\ell}'' + \ell(\ell+1)r^{-2} \psi_{\ell} + V(r) \psi_{\ell} = E \psi_{\ell}$$

which is an ordinary linear homogeneous second-order differential equation with a regular singular point at $r=0$. (It is assumed that

$$\lim_{r \rightarrow 0} r^2 V(r) = 0$$

Otherwise, this statement would not necessarily be true.) There are two classes of solutions: The regular solutions and the irregular solutions (all with respect to the point $r=0$). A regular solution is uniquely determined by a boundary condition at the origin—for example,

$$\lim_{r \rightarrow 0} r^{-\ell-1} \phi_{\ell}(r) (2\ell+1)!! = 1$$

For $\ell=0$ (the so-called "s wave") to which we shall for simplicity from now on restrict ourselves, this means

$$\phi(0) = 0 \quad \phi'(0) = 1$$

Every other regular solution is a multiple of this one. This includes the physical radial wave function. The asymptotic behavior of a regular solution for large r is given by

$$\phi(r) \simeq A \sin(kr + \delta)$$

where δ is the (s wave) phase shift.

In this context, then, the inverse scattering problem is posed by assuming the asymptotic phase $\delta(E)$ of $\phi(r)$ given for all positive E , to determine $V(r)$ in the radial Schrödinger equation. Mathematically, therefore, it is an inverse Sturm-Liouville problem on a semi-infinite interval $0 \leq r < \infty$. We are given asymptotic information about the regular solution, as a function of the spectral parameter, and we are asking for the reconstruction of the differential equation (in a specifically restricted class).

The attack on this problem involves two essential steps. As solutions of a Sturm-Liouville problem on a semi-infinite interval the functions $\phi(E, r)$ form a complete set. This statement may be written in the symbolic form of a Stieltjes integral

$$\int_{-\infty}^{\infty} d\rho(E) \phi(E, r) \phi(E, r') = \delta(r - r')$$

The spectral function $\rho(E)$ has a finite number of discontinuities at the discrete (point) eigenvalues of the Sturm-Liouville problem (at negative values of E) and it is continuously increasing for $E > 0$. The point eigenvalues are the bound states of the system, and if

$$\int_0^{\infty} dr r |V| < \infty$$

then their number is finite. Furthermore they cannot be positive. The positive spectrum is absolutely continuous. The two steps in the inverse problem are (1) $\delta(E) \rightarrow \rho(E)$, and (2) $\rho(E) \rightarrow V(r)$.

Step 1: $\delta(E) \rightarrow \rho(E)$

Let us define $\chi(E, r)$ to be a regular solution of the radial Schrödinger equation that satisfies the scattering integral equation

$$\chi = \chi_0 + G_0^+ V \chi$$

where

$$\chi_0(E) = k^{-1} \sin kr$$

and G_0^+ is the resolvent of the free Hamiltonian operator on the upper rim of its cut,

$$G_0^+(F) = (E + i0 - H_0)^{-1} \quad H_0 = \frac{-d^2}{dr^2}$$

Its kernel, the free Green's function, can be written down explicitly,

$$G_0^+(E; r, r') = -k^{-1} \sin kr < e^{ikr} >$$

where $r_< = \min(r, r')$ and $r_> = \max(r, r')$. For the bound states E_n we will assume χ to be normalized to unity:

$$\int_0^\infty dr |\chi_n(r)|^2 = 1$$

Now, since χ and ϕ are both regular solutions of the radial Schrödinger equation, they must be linearly dependent: $\phi(r) = f\chi(r)$. From the integral equation for χ one easily finds that as $r \rightarrow 0$

$$\chi(r) \sim r \left[1 - \int_0^\infty dr' e^{ikr'} V(r') \chi(r') \right]$$

comparison of which with the boundary condition for $\phi(r)$ shows that

$$f = 1 + \int_0^\infty dr e^{ikr} V(r) \phi(r)$$

On the other hand, as $r \rightarrow \infty$ we find from the integral equation

$$\chi(r) \sim \frac{1}{2ik} \left\{ \left[1 - 2i \int_0^\infty dr' \sin kr' V(r') \chi(r') \right] e^{ikr} - e^{-ikr} \right\}$$

and consequently, as $r \rightarrow \infty$

$$\chi(r) \sim \frac{i}{2k} \left(e^{-ikr} - \frac{f^*}{f} e^{ikr} \right)$$

because $\phi(r)$, being the solution of a real differential equation with real boundary conditions, is real (for real k). It follows that $f^*/f = e^{2i\delta}$ — that is, $\arg f = -\delta$. The function $f(E)$ is called the Jost function and it can be shown to be equal to the Fredholm determinant of the integral equation for χ .

On the other hand, the completeness of the spectrum of the Sturm-Liouville operator $H = -d^2/dr^2 + V(r)$ is expressed in terms of χ as follows:

$$\sum_n \chi_n(r) \chi_n(r') + \frac{1}{\pi} \int_0^\infty dE k \chi(E, r) \chi^*(E, r') = \delta(r - r')$$

if E_n are the point eigenvalues of H . Comparison with the spectral resolution in terms of ϕ shows that therefore

$$\frac{d\rho(E)}{dE} = \begin{cases} \frac{k}{\pi} \frac{1}{|f(E)|^2} & E > 0 \\ \sum_{n=1}^N \frac{\delta(E - E_n)}{N_n^2} & E < 0 \end{cases}$$

where

$$N_n^2 = \int_0^\infty dr [\phi(E_n, r)]^2$$

Thus, the connecting link between the spectral function $\rho(E)$ and the asymptotic phase $\delta(E)$ is the Fredholm determinant (or Jost function) f . The crucial property of $f(E)$ that allows us to determine $|f(E)|$ from a knowledge of its phase $-\delta(E)$ is its analyticity and asymptotic behavior.

Since the resolvent $G_0(E)$ of H_0 is an analytic (operator-valued) function of E regular everywhere in the E plane cut from 0 to ∞ and tends to zero there at infinity, $f(E)$ is an analytic function of E regular in the cut E plane and tends to 1 there at infinity. Furthermore, $f(E)$ has the reality property

$$f^- = f(E - i0) = f^*(E + i0)$$

Therefore, if f has no zeros, its logarithm satisfies a "dispersion relation"

$$\log |f(E)| = \frac{1}{\pi} P \int_0^\infty \frac{dE' \delta(E')}{E - E'}$$

(i.e., a Hilbert-transform relation between its real and imaginary parts) where P denotes Cauchy's principal value. If, on the other hand, $f(E)$ has zeros in the complex plane, then these are the point eigenvalues E_n of H . In that case we form the function

$$\bar{f}(E) \equiv f(E) \prod \frac{k - k_n^*}{k - k_n}$$

which has no zeros and hence satisfies the above Hilbert transform relation. We are therefore free to prescribe independently both the bound-state energies E_n , and the phase shift $\delta(E)$, and we can calculate $|f(E)|$ explicitly.

Thus the positive energy part of the spectral function $\rho(E)$ is determined uniquely by the phase shift $\delta(E)$ and the bound states together. The negative-energy part then still contains a positive normalization parameter for each bound state. Thus, prescription of $\delta(E)$ and of the N bound-states determines an N -parameter family of spectral functions.

Step 2: $\rho(E) \rightarrow V(r)$

We now turn to the second step, from $\rho(E)$ to V . By a method similar to the proof of the completeness of the functions ϕ , one can show that if $V^{(1)}$ is an arbitrary comparison potential and $\phi^{(1)}$ and $\rho^{(1)}$ are the corresponding regular solution and spectral function, then there is an integral representation of $\phi(E, r)$ of the form

$$\phi(E, r) = \phi^{(1)}(E, r) + \int_0^r dr' K(r, r') \phi^{(1)}(E, r')$$

where

$$K(r, r') = \int dh(E) \phi(E, r) \phi^{(1)}(E, r')$$

with

$$h(E) = \rho^{(1)}(E) - \rho(E)$$

It is easily seen that the kernel $K(r, r')$ satisfies the partial differential equation

$$\left[\frac{\partial^2}{\partial r^2} - V(r) \right] K(r, r') = \left[\frac{\partial^2}{\partial r'^2} - V^{(1)}(r') \right] K(r, r')$$

One then verifies by means of this equation, the Schrödinger equations for ϕ and $\phi^{(1)}$, and of the above integral representation, that $K(r, r')$ must satisfy the boundary conditions

$$2 \frac{d}{dr} K(r, r) = V(r) - V^{(1)}(r) \quad K(r, 0) = 0$$

Next one inserts the integral representation for ϕ in the definition of $K(r, r')$, and one obtains a linear integral equation for $K(r, r')$,

$$K(r, r') = g(r, r') + \int_0^r dr'' K(r, r'') g(r'', r')$$

where

$$g(r, r') = \int dh(E) \phi^{(1)}(E, r) \phi^{(1)}(E, r')$$

This is the *Gelfand-Levitan equation*. It can be shown under very general conditions to have a unique solution. Thus the step from ρ to V is complete, as follows.

Adopting an arbitrary comparison potential $V^{(1)}$ (for example, $V^{(1)} = 0$), we calculate $g(r, r')$ from the given $\rho(E)$. Then the Gelfand-Levitan equation must be solved, and from $K(r, r)$ we obtain $V(r)$ uniquely.

We summarize the inversion procedure schematically in the following way:

$$\begin{array}{cccc} (1) & (2) & (3) & (4) \\ \frac{d\sigma}{d\Omega} \rightarrow A(E, \cos \theta) \rightarrow \delta_\ell(E) \rightarrow \rho_\ell(E) \rightarrow V(r) \end{array}$$

In the first step existence and uniqueness are assured so long as the cross section is smooth and small enough. Step (2) is straightforward and unique. In step (3) existence is assured by mild conditions on the functions $\delta_\ell(E)$, and it is unique if there are no bound states of angular momentum ℓ . Otherwise, the binding energies can be assigned independently, and there is a "phase and binding-energy equivalent" family of spectral functions $\rho_\ell(E)$, with as many parameters as bound states. Step (4) is unique, under mild restrictions, and existence is assured.

The final question that arises is whether we are assured that the cycle is always closed. In other words: We start with a certain class 1 of potentials and prove certain properties of the phase shifts—say, that they are in class 2. If we now start with any phase-shift function $\delta_\ell(E)$ in class 2, the inversion procedure leads to potentials in some class 1'. We would call the procedure closed if the class 2 were known for which $1' \subset 1$. It would be optimally closed if the class 1 were known for which 2 is such that $1' = 1$. The answers to these questions are not known.

APPENDIX

The first attempts at a solution of the inversion problem here discussed were made by *Fröberg* [1947-51] and *Hylleraas* [1948]. It was then shown by *Bargmann* [1949] that the phase shift alone does not necessarily uniquely determine the potential. The uniqueness question was resolved by *Levinson* [1949], *Marchenko* [1950], and *Borg* [1949] (see also *Borg* [1946]). The previous work for *Fröberg* and *Hylleraas* was amended in the light of the new knowledge by *Holmberg* [1952]. New construction procedures were then invented by *Jost and Kohn* [1952]. For motivation see also *Jost* [1947].

Meanwhile the Gelfand-Levitan equation had been introduced into the inverse Sturm-Liouville problem by *Gelfand and Levitan* [1951]. It was then applied to the inverse-scattering problem by *Jost and Kohn* [1953] and *Levinson* [1953]. Since that time the problem has given rise to a rather extensive Russian mathematical literature: *Chudov* [1949 and 1956]; *Krein* [1951-58]; *Marchenko* [1952-60]; *Berezanskii* [1953-58]; *Block* [1953]; *Firsov* [1953]; *Stashevskaya* [1953]; *Volk* [1953]; *Neigauz* [1955]; *Levitan* [1956-64]; *Faddeev* [1956-64]; *Agranovich and Marchenko* [1957-63]; *Gohberg and Krein* [1958]; *Gasymov* [1963]; *Kostarev* [1964]; *Lavrentev* [1964 and 1956]; *Levitan and Gasymov* [1964]; *Kac* [1965]; *Gugushvili* [1970]. See also *Ambarzumian* [1929]; *Jauho* [1950]; *Crum* [1955]; *Kay* [1955 and 1961]; *Ohmura* [1956]; *Jost* [1956]; *Fulton* [1956]; *Friedman* [1957]; *Tierz* [1959]; *Levitan and Sarkisyan* [1960]; *Martin* [1961]; *Petras* [1962]; *Blazek* [1962-66]; *Hylleraas* [1964]; *Pearce* [1964]; *Ramm* [1964]; *Ljance* [1964 and 1966]; *Swan and Pearce* [1966]; *Lax and Phillips* [1966]; *Portinari* [1966 and 1967]; *Prosser* [1969]; *O'Brien and Bernstein* [1969]; *Newton* [1966 and 1970].

A more general approach started with the inversion of the electromagnetic reflection problem and led to application in the full three-dimensional Schrödinger equation: *Kay and Moses* [1955-61]; *Moses* [1956]; see also *Faddeev* [1966].

Generalizations of the Gelfand-Levitan and Marchenko procedures to coupled equations were given by *Newton and Jost* [1955]; and *Krein* [1956]; to the scattering of spin 1/2 particles with a tensor force present by *Newton* [1955]; *Agranovich and Marchenko* [1958]; *Fulton and Newton* [1956]; *Newton and Fulton* [1957]; (see also sec. 9 of *Newton* [1960]); to the Klein-Gordon equation by *Corindaldesi* [1954]; to the Dirac equation by *Prats and Toll* [1959]; *Gasymov and Levitan* [1966]; and *Gasymov* [1968]; to both relativistic equations by *Verde* [1958]; to coupled channels by *Cox* [1962]; to separable potentials by *Gourdin and Martin* [1957 and 1958], *Chadan* [1958 and 1967]; *Bolsterli and Mackenzie* [1965]; *Mills and Reading* [1969]; *Fiedeldey* [1969]; and *Tabak* [1969]; to energy-dependent potentials by *Mal'cenko* [1966]; and to charged particles by *Swan* [1967].

For some applications, see *Newton* [1957]; *Gardner et al.*, [1967]; *Buck* [1969]; *Fiedeldey* [1970]; *Fuda* [1970]; *Srivastava* [1970]; *Child* [1970].

DISCUSSION

Unidentified speaker: This development is a very pretty theoretical one. This inverse problem demands that you know the spectral weight function for all values of the energy, but when you measure it in the physical laboratory you are bounded by practical situations. How stable is this process against uncertainties or incompleteness in the data?

Newton: Not much can be said about that. Of course, you know the data only up to a finite energy and then not exactly. There are some things that can be said about the stability of the process with respect to small perturbations at a finite energy, but the main limitation of the method is that you have to know the phase shift up to infinite energy.

Unidentified speaker: Suppose you know it only to a finite energy, what class of potential can you fit by a class of phase shifts that is given only up to a finite energy?

Newton: That's a difficult question and not much is known. What you'd like to know are certain specific features of the potential. How does it behave asymptotically? Can you tell whether the potential goes down exponentially for large distance? If so, how? The answer there is discouraging. If the potential goes down exponentially, then the S matrix is an analytical function in a strip around the real axis. Suppose for a very large energy, I change the phase shift by a little kink—that is, I give its first derivative a discontinuity—then the S matrix cannot possibly be analytic in that strip any longer and therefore the potential cannot go down exponentially.

Moses: Dr. S. C. Wang, MIT, tried some numerical experiments on the stability of the Gelfand-Levitan equation. He varied the normalization constants a little. It is quite stable. I also tried to make nonanalytic scatter operators, jumps, step functions, etc. I didn't solve them although I think it is solvable. I think you will get unitary scattering operators and so on, but I think you get singular potentials in the finite domain.

Parker: Even in the ideal case, you still have an n -parameter ambiguity, a nonuniqueness—the number of bound states. In the case where $Q > 1$, what is the parameterization? Is it countable, or nondenumerable, or how bad is the nonuniqueness when you've got counter examples?

Newton: We're talking about two different things. You're talking about the inverse problem from the cross section to the amplitude. That has nothing to do with the number of parameters. In the example that's known, it's a threefold ambiguity. It was generated numerically. Somebody took a cross section and searched for different amplitudes to fit it. He found three solutions on a computer.

Moses: John Lamont and I looked at this problem, but we found that if you did a measurement through slits you can find complex amplitudes from the corresponding cross sections.

Newton: There is also another scattering experiment in principle, to determine the phase of the amplitude, based on the Hanbury-Brown and Twiss effect of intensity interferometry. Whether this is feasible in the particle domain is doubtful.

Puinney: I am concerned with the existence of states that are bound at positive energies and can tunnel out. I imagine these are inaccessible to experimental data with experimental errors. How are these viewed practically?

Newton: There are two kinds of such states. One is simply a resonance, a narrow peak in the cross-section, which can be handled by the inversion method. The other thing is actual bound states at positive energy. For elastic scattering by local potentials that is impossible. For nonlocal potentials, or a three-body problem, this is possible. They may cause difficulties.

Phinney: I am concerned with very broad resonances. I have a seismological analog concerning seismic body waves that are observed at a variety of distances at a given frequency. These signals illuminate other regions of the earth but are not permitted to enter the earth's core unless tunneling is permitted. From a geometrical optics view there is no penetration. If the core had a sufficiently low velocity that the states were perfectly bound and the core actually had resonances that could not be observed, I would see where the nonuniqueness lay. We are unable to illuminate these regions of the core, and yet the states that exist in the core are essentially resonances and the effects of the core are down in the noise.

Newton: The difference is that you don't ask for specific mechanisms. You are simply taking the phase shift as a function of the energy for granted. Your problem is much more complicated.

Unidentified speaker: I would like to make a comment on the application to seismology. The seismological problem was modeled at Lincoln Labs some years ago as a one-dimensional, layered model. In that case, the inverse problem is like the problem of a transmission line in which the impedance, or index of refraction is discontinuous. The kinds of effects that you look for, leaky states, etc., can be taken into account by this same algorithm but with a different class of potentials.

Unidentified speaker: Yesterday we heard problems concerning a complex index of refraction, which would correspond to a complex potential here, including absorptive effects. Can you generalize your procedure to handle that?

Newton: Yes.

REVIEW OF THE INVERSE SCATTERING PROBLEM AT FIXED ENERGY IN QUANTUM MECHANICS

Pierre C. Sabatier

Département de Physique Mathématique*

Université des Sciences et Techniques de Montpellier

N73-11609

ABSTRACT

A short review is given of the methods of solution of the inverse scattering problem at fixed energy in quantum mechanics. The attempt has been made to cover the literature up to April 1971.

INTRODUCTION

We are interested in scattering experiments of a beam of particles A at a *nonrelativistic* energy E by a target made of particles B . For the quantum mechanical description of the system $A - B$, we use the Schrödinger equation and one or several functions depending on the relative distance of the particles. Our *inverse problem* is the construction of these *potentials* from the experimental measurements.

In the simplest case—namely, for spinless particles and isotropic potentials—the system may be described in its own center-of-mass frame by the Schrödinger equation:

$$\Delta \Psi(x) + [1 - V(x)] \Psi(x) = 0 \quad (1)$$

The above equation can be separated; hence, we are led to the so-called *radial*, or partial wave, equation:

$$D_r^V \psi_\ell(r) = \ell(\ell + 1) \psi_\ell(r) \quad (2)$$

where

$$D_r^V = r^2 \left[\frac{\partial^2}{\partial r^2} + 1 - V(r) \right] \quad (3)$$

The partial wave $\psi_\ell(r)$ is the solution of Eq. (2) whose principal part, as r goes to zero, is $(1/2r)^{\ell+1} \Gamma(1/2)/\Gamma(\ell+3/2)$, so that for $V = 0$ we obtain:

*Physique Mathématique et Théorique, Equipe de recherche associée au C.N.R.S.

$$\psi_{\ell}^0(r) \equiv u_{\ell}(r) = \left(\frac{1}{2} \pi r\right)^{1/2} J_{\ell+1/2}(r) \quad (4)$$

For a *physical* potential, the function $\psi_{\ell}(r)$ must have the following asymptotic behavior:

$$\psi_{\ell}(r) = A_{\ell} \sin \left(r - \frac{1}{2} \pi \ell + \delta_{\ell} \right) + o(1) \quad r \rightarrow \infty \quad (5)$$

A sufficient condition for the existence of the δ_{ℓ} is that the following quantity is finite:

$$\|V\| = \int_0^a \rho^{1-\epsilon} |V(\rho)| d\rho + \int_a^{\infty} |V(\rho)| d\rho \quad (6)$$

where a is a general length.¹ This condition defines a set \mathcal{O} of potentials, in which $\|V\|$ is obviously a norm. However, the set \mathcal{P} for which the δ_{ℓ} are defined by Eq. (5) is much larger than \mathcal{O} : \mathcal{P} contains, for instance, infinitely repulsive potentials. The cross section is related to the phase shifts by

$$f(\theta) = \sum_{\ell=0}^{\infty} (2\ell+1) e^{i\delta_{\ell}} \sin \delta_{\ell} P_{\ell}(\cos \theta) \quad (7)$$

$$\sigma(\theta) = |f(\theta)|^2 \quad (8)$$

The scattering amplitude $f(\theta)$ appears in the asymptotic behavior of the wave function

$$\psi(r) = e^{ik^{-1}k \cdot r} - f(\theta) r^{-1} e^{ir} \quad (9)$$

The conditions for which Eqs. (7) and (9) make sense are not necessarily met for any potential of \mathcal{P} . Because of the bounds of the Legendre polynomials for $\sin \theta \neq 0$, a sufficient condition for the convergence of Eq. (7) is that $|\delta_{\ell}|$ be bounded by $C\ell^{-3/2-\epsilon}$. Now, according to a result of *Martin* [1965], the phase shifts are bounded for large ℓ by

$$C \int_0^{\infty} u_{\ell}^2(\rho) |V(\rho)| d\rho$$

provided that this quantity goes to zero. Therefore, a *sufficient* condition for a good definition of $f(\theta)$ is that $\rho^{5/2+\epsilon} |V(\rho)|$ be bounded for large ρ .

¹ Throughout the paper, C is a general constant, ϵ an arbitrarily small positive number, a a general length. They are not intended to have the same value every time they are used. The term "bounded" is often used in place of "absolutely bounded."

This, however is much too strong a condition in general. Let us define \mathcal{P}' as the set of potentials for which Eqs. (7) and (9) make sense; \mathcal{P}' is the set in which one must look for the solutions of the inverse problem. In this simple case, the inverse problem can be formulated as follows: Construct the potentials $V(r)$, ($V \in \mathcal{P}'$) that yield the cross section $\sigma(\theta)$.

As in most inverse problems, the questions of interest changed with time throughout the successive studies of this problem. In the first studies, authors tried to obtain an approximate solution of the problem. The quest for exact solutions led to the realization of the nonuniqueness of solutions and its consequences, which suggested further questions. There are six main questions essential in the inverse problem at fixed energy.

1. Given $\sigma(\theta)$, does a potential V exist in \mathcal{P}' that generates σ at the energy E through the Schrödinger equation? Such a potential is called a *solution* of the inverse problem at the energy E .
2. Let us assume a class \mathcal{w} of functions ($\mathcal{w} \subseteq \mathcal{P}'$) that contains a solution V of the inverse problem at the energy E . Is V the unique solution in \mathcal{w} ?

The answer to (1) being in general affirmative and the answer to (2) being in general negative, we are led to:

3. Give a method of constructing a solution of the problem.
4. Let us call "equivalent" two potentials that yield the same cross section. Give methods for constructing all the equivalent potentials in \mathcal{P}' , or in a large enough, well-defined subset of \mathcal{P}' .
5. Give an appraisal of the deviation from each other of all the equivalent potentials in \mathcal{P}' or in a large enough, well defined, subset of \mathcal{P}' .

In (1) through (5), it is assumed that the cross section was exactly known. The experimental scattering data are necessarily affected by errors. We must find out whether the inversion procedures are stable with respect to the perturbation generated by the noise of the data:

6. Describe the evolution of the set of equivalent potentials when the cross section is submitted to random perturbations.

The above are strictly concerned with the inverse problem at fixed energy. In a more general framework, an interesting question would be: Let $\sigma(\theta)$ be given as a function of E . What conditions must be fulfilled for which at least one of the solutions is a static potential, and how do we obtain that solution?

We have formulated all our questions for the simple problem of a spherically symmetric spinless potential. With trivial modifications, they can be extended to the various generalizations of the problem.

The historical development of the subject has been similar to that of other inverse problems. The early studies tried to solve questions (1) and (3) by trial and error, or by approximate methods. General methods began to appear in 1959, but they remained widely formal until a few years ago when they were systematically developed, giving definite answers to (1), (2), and (3). The solutions, or attempts at solution, to (4), (5), and (6) are very recent, and the subject is far from being exhausted. On the whole, the evolution of the inverse scattering problem at fixed energy has been very rapid, and about 30 papers have been sufficient to achieve a state of solution equivalent to that requiring about 100 papers for the inverse problem at fixed angular momentum. One reason for this rapid evolution, of course, is that many tools could be constructed by analogy with the inverse problem at fixed ℓ . There are, however, several tools introduced in the present problem, such as the interpolation coefficients, or the scattering structure function, that were developed independently of the inverse problem at fixed ℓ , and their applicability to that problem would be an interesting subject for study.

Here we follow the historical order in the evolution of the inverse scattering problem at fixed energy. Except for trial-and-error methods, the published studies gave separate solutions to the construction of the scattering amplitude from the cross section and to the construction of the potential from the scattering amplitude.

TRIAL-AND-ERROR METHOD

The trial-and-error method certainly is the simplest way to give an answer to question (3). Calculations of cross sections are done for a series of trial potentials until a satisfactory fit is obtained. Hence, the following answer to (3) can be obtained: Here is a potential that yields a cross section approximately equal to the experimental cross section.

Because of the experimental errors, three parameters related to the depth of the potential, its range, and the range of its "surface," respectively, prove to be sufficient to define the class of trial potentials. The most popular shape of potential including these parameters seems to be the Saxon-Woods shape:

$$V(r) = A \left[\frac{1 + \exp(r - R)}{a} \right]^{-1} \quad (10)$$

where A is the potential depth, R its range, and a the surface width.

The method can easily be generalized. It has been used for potentials including spin-orbit forces, tensor forces, imaginary parts, and more generally all situations in which the direct problem has been solved. Fitting procedures have been made as reliable and efficient as possible.

Any trial-and-error "method" is nothing more than an algorithm in which a series of observable quantities is calculated from the interaction and compared with the experimental result. Therefore, the differences between the various procedures are concerned either with numerical analysis or with computer tricks. The structure of the method, its physical interest, and its fundamental limits are always the same. Furthermore, the *only* information that can be obtained by a trial-and-error method for the inverse problem is the partial answer to (3) quoted above.

On the other hand, the trial-and-error methods are the only ones in which the existence of a static potential consistent with the experimental results has been studied. In nuclear physics, where extensive studies have been done, trial-and-error methods have made plausible the proposal that the nuclear potential depends smoothly on the energy, and since this dependence is also predicted by the theory of nuclei, we have every reason to believe it. Also, the failure of trial-and-error methods to yield a static potential to fit the cross sections at various energies does not imply that such a potential does not exist. Results obtained by trial-and-error method are necessarily biased by the set of trial potentials used.

Textbooks, such as *Hodgson* [1963] contain good chapters on the fitting procedures used to apply the optical model in nuclear physics. Computer programs are available in most computer libraries. A classical example of a detailed description of a program of this kind has been published by *Melkanoff et al.* [1961].

Although we have given only a cursory appraisal of trial-and-error methods, their importance should not be underestimated. We believe, they are relevant to the *direct* problem—not to the *inverse* problem. They have been of significant value in the development of scattering studies in physics, and their success has been the best proof of the validity of the optical model in nuclear physics. Furthermore, the earliest recognition of the fundamental ambiguities of the problem can be attributed to trial-and-error studies.

CONSTRUCTION OF THE SCATTERING AMPLITUDE FROM THE CROSS SECTION

The problem of constructing the scattering amplitude from the exact knowledge of the differential cross section at a given energy could theoretically be avoided by performing special measurements on the scattered particles. *Goldberger et al.* [1963] noted this possibility on the basis of an idea proposed by *Hanbury, Brown and Twiss* [1956]. See also *Twiss et al.* [1957]; *Goldberger and Watson* [1964; 1965a,b] and *Goldberger et al.* [1966]. Although this suggestion may be of use in some cases, in particle physics it is impossible to obtain the correlated counting rates of two detectors with the beam intensities actually available. Furthermore, it is not obvious that by overcoming this difficulty we will be able to obtain a solution of the problem. In typical situations, the scattering phase remains deeply involved in the formulas. Therefore, mathematical ways of constructing the scattering amplitude have a good chance to remain useful for many years. The solutions of this first step of the inverse problem are reviewed by Newton in the preceding discussion.

CONSTRUCTION OF THE POTENTIAL FROM THE PHASE SHIFTS

Let us assume that the first step of the problem has given an infinite set of phase shifts. An analysis of the second step of the inverse problem must deal with the six fundamental questions given in the introduction. The only modification comes in questions (1) and (6), which we replace by:

- 1a. Given a sequence $\delta = \{\delta_\ell\}$ of phase shifts, consistent with Eqs. (7) and (9), does a potential V exist in \mathcal{P}' that generates δ at the energy E ? Such a potential is called a solution of the inverse problem at the energy E .
- 6a. Describe the evolution of the set of equivalent potentials when the cross section is submitted to random perturbations.

JWKB Method

Development— The method seems to have been introduced by *Firsov* [1953], and *Wheeler* [1955]. The presentation below is due to *Sabatier* [1966a], who gave arguments for the "right choice" of an interpolation. Extension to higher orders of the asymptotic approximation of the phase shifts also has been given by *Sabatier* and rediscovered by *Vollmer* [1969], who made numerical calculations using the method for practical inversion purposes. *Miller* [1969] has discussed the method in terms of its applications to problems in chemical physics. Note that the first-order method is equivalent to the inversion method in classical mechanics given by *Keller et al.* [1956].

The first attempt toward a discussion of this problem used the JWKB approximation. If we assume the validity of this approximation, the phase shift δ_ℓ is given by:

$$\delta_\ell = \lim_{A \rightarrow \infty} \left[\int_{r_\ell}^A \left\{ r^2 [1 - V(r)] - \left(\ell + \frac{1}{2} \right)^2 \right\}^{1/2} r^{-1} dr - \int_{\ell+1/2}^A \left[r^2 - \left(\ell + \frac{1}{2} \right)^2 \right]^{1/2} r^{-1} dr \right] \quad (11)$$

where the turning point r_ℓ is the largest zero of $\{ r^2 [1 - V(r)] - (\ell + 1/2)^2 \}$. Let us now introduce the function

$$\lambda(r) = r[1 - V(r)]^{1/2} \quad (12)$$

Let us also assume that $\lambda(r)$ is equal to zero for $r = r_0$ and is a monotonically increasing function of r for $r > r_0$. Let $r = \psi(\lambda)$ be the inverse function of $\lambda(r)$. From Eq. (11) it follows that:

$$\delta_\ell = \int_{\ell+1/2}^{\infty} \left[\lambda^2 - \left(\ell + \frac{1}{2} \right)^2 \right]^{-1/2} H(\lambda) \lambda d\lambda \quad (13)$$

where

$$H(\lambda) = \log \frac{\lambda}{\psi(\lambda)} \quad (14)$$

Within the above assumptions, $V(r)$ can be derived readily from $H(\lambda)$. The problem reduces to the determination of $H(\lambda)$ from the δ_ℓ through Eq. (13). This is straightforward if δ_ℓ is known as a differentiable function of $(\ell + 1/2)$, say $\Phi(\ell + 1/2)$, going to zero more rapidly than $(\ell + 1/2)^{-1-\epsilon}$ as ℓ goes to ∞ . Then

$$H(\lambda) = -2\pi^{-1} \int_{\lambda}^{\infty} \Phi'(x)(x^2 - \lambda^2)^{-1/2} dx \quad (15)$$

Applications— Suppose we are given an infinite sequence $\{\delta_\ell\}$. The first a priori assumption is that the function $\lambda(r)$ corresponding to the *unknown* potential is monotonically increasing. Physically, this means that there must be no orbiting. Then we have to decide what interpolation we have to choose. Now, from Eq. (7) we derive

$$\exp [2i\delta_\ell] = 1 + i \int_0^\pi f(\theta) P_\ell(\cos \theta) \sin \theta d\theta \quad (16)$$

One should not conclude from Eq. (16) that the "natural" interpolation of δ_ℓ corresponds to the usual interpolation of $P_\ell(\cos \theta)$. Rather, an analysis of the errors due to the JWKB approximation for an even potential shows that the best interpolation may correspond to an unusual interpolation of $P_\ell(\cos \theta)$ [Sabatier, 1966a]. The existence of an infinity of interpolations shows that the apparent uniqueness of the inversion procedure is false.

Actually, we see below that the choice of an interpolation completely determines the exact potential. On the other hand, only smooth interpolations are consistent with the assumptions of the JWKB approximation, so that it is reasonable to use any smooth continuous fitting of the phase shifts, and to take for granted, or to check a posteriori, that the various potentials which may be obtained are not very different from each other. This way of using the above method may be of interest for a real potential. When it is used for a complex potential, the experimental errors on the strongly absorbed $\exp [2i\delta_\ell]$ can yield enormous errors on the potential [Sabatier, 1966a]. Therefore, we conclude that the method above is not really an inversion method but a fitting procedure. It yields to the inverse problem the answer given by trial-and-error methods, nothing more. It may be of interest to save computer time or to study qualitatively the relevance of a model.

A Special Approach for Superposition of Yukawa Potentials

Martin and Targonski [1961] looked for a solution of the problem in the class of potentials of the form

$$V(r) = r^{-1} \int_{\mu}^{\infty} C(\alpha) \exp[-\alpha r] d\alpha \quad (17)$$

where μ is positive, and $C(\alpha) \in L_1(\mu, \infty)$. Now let t be the momentum transfer [$t = -2(1 - \cos \theta)$]. Its physical values belong to $[-4, 0]$. Let $T(t)$ be the scattering amplitude. If $V(r)$ is given by Eq. (17), $T(t)$ is analytic in the complex t plane, except along a cut on the real axis from $t = \mu^2$ to $t = +\infty$, and the following dispersion formula, or a conveniently subtracted form, holds:

$$T(t) = \int_{\mu^2}^{\infty} D(t')(t' - t)^{-1} dt' \quad (18)$$

where $2\pi i D(t)$ is the discontinuity across the cut. This formula can also be considered as an integral equation for $D(t)$. It enables one to obtain $D(t)$ by extrapolating the values that $T(t)$ assumes in the physical region. Once the discontinuity has been obtained, the potential can be reconstructed. More precisely, *Martin and Targonski* [1961] show that in the Born expansion of $T(t)$, the discontinuity across the cut in the region $\mu^2 < t < (n+1)^2 \mu^2$ comes from the n first Born terms only. Let us now introduce the expansion

$$V(r) = \sum_1^n V_\ell + V_{n+1} \quad (19)$$

where

$$\begin{aligned} V_\ell &= r^{-1} \int_{\ell\mu}^{(\ell+1)\mu} C(\alpha) \exp[-\alpha r] d\alpha \quad \ell \leq n \\ V_{n+1} &= r^{-1} \int_{(n+1)\mu}^{\infty} C(\alpha) \exp[-\alpha r] d\alpha \end{aligned} \quad (20)$$

note that V_{n+1} does not contribute to the discontinuity for $\mu^2 < t < (n+1)^2 \mu^2$, while V_n does contribute through T_1 only. Now, assume that $C(\alpha)$ is known from $\alpha = \mu$ to $\alpha = n\mu$. This yields V_1, V_2, \dots, V_{n-1} , from which the discontinuity of $T_2 + \dots + T_n$ in the region $n^2 \mu^2 < t < (n+1)^2 \mu^2$ can be derived. We can then obtain $C(\alpha)$ in this interval through the formula

$$2\pi i t^{-1/2} C(t^{1/2}) = 2\pi i D(t) - \text{discontinuity of } (T_2 + \dots + T_n)$$

Through this iterative procedure, we can reconstruct $C(\alpha)$ from the discontinuity of T across the cut. *Martin and Targonski* [1961] give the necessary (but not necessarily sufficient) conditions for the scattering amplitude to be generated by a superposition of Yukawa potentials. The uniqueness of the solution follows from the analytic continuation and the iteration procedure. The method can be easily extended to exchange forces. It can also probably be extended to potentials bounded by decreasing exponentials (*Martin*, private communication).

Numerical computations are not feasible. Besides, since the method involves an analytic continuation, it is generally unstable with respect to experimental or roundoff errors [Viano, 1969].

Approaches Through a Gelfand-Levitan Procedure

Regge [1959] introduced an approach to the scattering problems at fixed energy fairly close to the Gelfand-Levitan procedure introduced in problems at fixed ℓ . In such a method, given two potentials V_0 and V , one looks for a "transformation kernel" $K_{V_0}^V(r, r')$, which generates the wave function ψ_ℓ^V from the wave function $\psi_\ell^{V_0}$ through

$$\psi_\ell^V(r) = \psi_\ell^{V_0}(r) - \int_0^r K_{V_0}^V(r, \rho) \psi_\ell^{V_0}(\rho) \rho^{-2} d\rho \quad (21)$$

$$K_{V_0}^V(r, r') = -\frac{1}{2} r \int_0^r \rho [V(\rho) - V_0(\rho)] d\rho \quad (22)$$

Therefore, if V_0 is a well-known potential with well-known wave functions, the scattering problem reduces to the determination of $K_{V_0}^V(r, r')$. For this, it is convenient to introduce an auxiliary tool, the input function $f_{V_0}^V(r, r')$, which is a solution of the partial differential equation

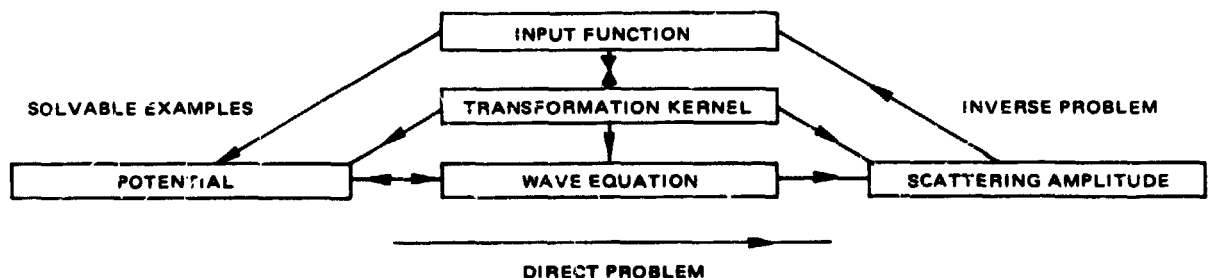
$$(D_r^{V_0} - D_{r'}^{V_0}) f_{V_0}^V(r, r') = 0 \quad (23)$$

$$f_{V_0}^V(r, 0) = f_{V_0}^V(0, r') = 0$$

and enables one to obtain the transformation kernel through the "fundamental" integral equation

$$K_{V_0}^V(r, r') = f_{V_0}^V(r, r') - \int_0^r K_{V_0}^V(r, \rho) f_{V_0}^V(\rho, r') \rho^{-2} d\rho \quad (24)$$

obtained by analogy with the Gelfand-Levitan equation. Therefore, the scattering problem is completely solved when $f_{V_0}^V(r, r')$ is known. A convenient representation of this machinery is given by the triangular diagram [Sabatier, 1968].



Note that Eq. (24) is a Fredholm equation. If $r' \rightarrow f_V^V(r, r')$ belongs to $L_2(0, r)$, the Fredholm alternative holds. If we keep only continuous solutions of Eq. (23), the values of r for which the homogeneous form of Eq. (24) has a solution correspond to singularities of $V(r)$. At any regular point of $V(r)$, therefore, Eq. (24) has one solution only. With weak additional assumptions, it is possible to show that the singularities are isolated. It is easy to show from Eqs. (23) and (24) that the following equations hold at any regular point:

$$\left(D_r^V - D_{r'}^{V_0} \right) K_{V_0}^V(r, r') = 0 \quad (25)$$

$$K_{V_0}^V(r, 0) = K_{V_0}^V(0, r') = 0$$

Two kinds of studies have used $f_V^V(r, r')$ as the fundamental tool for a solution. In the Regge-Loeffel method, the questions (1) and (2) are well studied, but (3), (4), and (5) are definitely neglected; the emphasis, therefore, is on a general representation of $f_V^V(r, r')$. In the Newton-Sabatier method, question (3) is dominant, and (1) and (2) follow; the emphasis, therefore, is on representations of $f_{V_0}^V(r, r')$ that permit effective construction of the potentials.

Regge-Loeffel Method – Introduced by Regge [1959] and thoroughly studied by Loeffel [1968], this method makes use of the above machinery with $V_0 = 1$. The unperturbed partial wave equation thus reads:

$$r^2 \frac{\partial^2}{\partial r^2} \psi_\ell^1(r) = \ell(\ell+1) \psi_\ell^1(r) \quad (26)$$

Hence, $\psi_\ell^1(r)$ reduces to a power of r . Using Gelfand-Levitan [1951] methods, Loeffel was able to prove that the function $f_1^V(r, r')$ corresponding to any potential of σ has the following representation

$$f_1^V(x, y) = -(2\pi)^{-1} \int_{-\infty}^{+\infty} \gamma(\tau)(xy)^{i\tau-1/2} + \sum_{k=1}^{\infty} d_k(xy)^{\nu_k-1/2} \quad (27)$$

where the continuous functions $\gamma(\tau)$ and the sequences d_k and ν_k are unique for a given potential. Actually, these quantities appear in the spectrum of the differential operator $-(d/dx)[x^2(d/dx)] - 1/4 + x^2[V(x) - 1]$ and thus are termed the *spectral data* by Loeffel. The series in Eq. (27) converges uniformly for (x, y) in any compact contained in $R_+ \times R_+$. The integral symbolizes a unitary operator U defined on $L_2(0, \infty)$ whose application to a function of $L_2(0, \infty)$ should take into account the convergence in L_2 of all the integrals. Using this complete characterization of any potential in σ , it has been possible to prove by function-theoretic methods the two following uniqueness theorems [Loeffel, 1968], where ν is the "continuous variable" $\ell + 1/2$:

Theorem 1. Let V_1 and V_2 be in σ . Let α_1 and α_2 be the corresponding Jost functions. If $\alpha_1(\nu) = \alpha_2(\nu)$ for $\text{Re } \nu \geq 0$, then $V_1(x) = V_2(x)$ for almost all positive x .

Theorem 2. Let $s_\ell = \exp[2i\delta_\ell]$, let $\sigma(\ell + 1/2)$ be the interpolation of s_ℓ obtained through the Jost functions (the so-called "Regge interpolation"), and let V_1 and V_2 be in the class \mathcal{O} . If the corresponding Regge interpolations σ_1 and σ_2 satisfy $\sigma_1(\nu) = \sigma_2(\nu)$ for all ν with $\text{Re } \nu > 0$ where both are holomorphic, then $V_1(x) = V_2(x)$ for almost all positive x .

The problem of uniqueness therefore reduces to the step $s_\ell \rightarrow \sigma(\ell + 1/2)$. At this point, there can be uniqueness only for particular classes of potentials that allow the interpolation to be unique. This holds if the analytic properties of the Jost functions enable one to apply a uniqueness theorem such as Carlson's theorem or the Lagrange-Valiron theorem. If a subclass of σ is chosen such that the solution is unique, and if the Regge interpolation is known, question (1) can be answered positively in certain cases [Loeffel, 1968; Burdet et al., 1965] and formal answers can be given to (3). However, attempts to answer (5) and (6) are lacking, and no constructive method is available. Besides, the stability of a method using an interpolation process as an intermediate step is highly questionable [Viano, 1969].

Newton-Sabatier methods — Let us use the machinery quoted above, with $V_0 = 0$, and following Newton [1962] let us introduce the series

$$f_0^V(r, r') = \sum_0^\infty c_\ell u_\ell(r) u_\ell(r') \quad (28)$$

where $|c_\ell| < C(1 + \ell)$ for any ℓ . Hence, the expansion

$$K_0^V(r, r') = \sum_0^\infty c_\ell \psi_\ell(r) u_\ell(r') \quad (29)$$

follows for $K_0^V(r, r')$. Substituting Eq. (29) in Eq. (21), we obtain

$$\psi_\ell^V(r) = u_\ell(r) - \sum_{\ell'} L_{\ell\ell'}(r) c_{\ell'} \psi_{\ell'}(r) \quad (30)$$

where

$$L_{\ell\ell'}(r) = \int_0^r u_{\ell'}(\rho) u_\ell(\rho) \rho^{-2} d\rho \quad (31)$$

The infinite system of coupled linear algebraic Eq. (30) is equivalent to the integral Eq. (24). Letting $r \rightarrow \infty$ in Eq. (30) and using Eq. (5) we obtain the system

$$A_\ell e^{i\delta_\ell} = 1 - \frac{1}{2} \pi \frac{c_\ell - A_\ell e^{i\delta_\ell}}{(2\ell + 1)} + i \sum_{\ell'=0}^\infty M_{\ell\ell'} c_{\ell'} A_{\ell'} e^{i\delta_{\ell'}} \quad (32)$$

where

$$M_{\ell\ell'} = \begin{cases} \left[\left(\ell' + \frac{1}{2} \right)^2 - \left(\ell + \frac{1}{2} \right)^2 \right]^{-1} & \text{if } \ell' - \ell \text{ is odd} \\ 0 & \text{if } \ell' - \ell \text{ is even} \end{cases} \quad (33)$$

Considering separately the real and the imaginary parts of Eq. (32), we obtain the systems:

$$\tan \delta_Q = \sum_{Q'} M_{QQ'} a_{Q'} (1 + \tan \delta_Q \tan \delta_{Q'}) \quad (34)$$

$$c_Q = a_Q (1 + \tan^2 \delta_Q) \left[1 - \frac{1}{2} \pi \frac{a_Q}{2l + i} (1 + \tan^2 \delta_Q) - \sum_{Q'} M_{QQ'} a_{Q'} (\tan \delta_{Q'} - \tan \delta_Q) \right]^{-1} \quad (35)$$

where

$$a_Q = c_Q A_Q \cos \delta_Q \quad (36)$$

In matrix notation, system in Eq. (34) reads

$$\tan \Delta e = M(1 + R)a \quad (37)$$

where

$$R = M^{-1} \tan \Delta M \tan \Delta \quad (38)$$

This ends the formal inversion procedure. It is clear that if M^{-1} and $(I + R)^{-1}$ can be obtained, a set of coefficients a_Q can be obtained from Eq. (37). When inserted in Eq. (35), they yield a set of $c_{Q'}$ that yield all the quantities in the problem. Although this method appears very simple, progress beyond these formal aspects has been surprisingly difficult.

Using the properties of an auxiliary matrix N , Newton [1962] was able to prove the existence of a matrix M^{-1} bisided inverse of M and of a vector v such that

$$Mv = 0 \quad (39)$$

A study by Redmond [1964] improved our knowledge of the matrix N , but not until 1966 were M^{-1} and v explicitly given, together with the analogous quantities corresponding to a class of infinite matrices [Sabatier, 1966b]. The result, for M is:

$$M^{-1} = -\mu M \mu \quad (40)$$

$$v_{2Q} = \mu_{2Q} \quad (41)$$

$$v_{2Q+1} = 0$$

where

$$\mu_{2n} = \frac{4}{\pi} \left[\frac{\Gamma(n+1/2)}{\Gamma(n+1)} \right]^2 \left(2n + \frac{1}{2} \right) \quad (42)$$

$$\mu_{2n+1} = \frac{4}{\pi} \left[\frac{\Gamma(n+3/2)}{\Gamma(n+1)} \right]^2 \left(2n + \frac{3}{2} \right) \quad (43)$$

When $(1 + R)$ is invertible, the solution of Eq. (37) is therefore

$$t = (1 + R)^{-1} (\alpha v + M^{-1} \tan \Delta e) \quad (44)$$

where α is an arbitrary parameter.

It has been possible [Sabatier, 1966b] to prove that the method is consistent for *almost* every set of phase shifts satisfying the inequality

$$|\tan \delta_\ell| < C(\ell+1)^{-3-\epsilon} \quad (45)$$

and that it yields, for each α , a potential $V(r)$ bounded for positive r and going to zero as

$$V(r) \sim C(\alpha - \beta) r^{-3/2} \cos \left(2r - \frac{1}{4}\pi \right) + O(r^{\epsilon-2}) \quad r \rightarrow \infty \quad (46)$$

where β depends on the phase shifts only. Among these equivalent potentials, we see that one, and only one, goes to zero more rapidly than $r^{-3/2}$ and may exhibit a nonoscillating tail. If the δ_ℓ are chosen equal to zero, a non-trivial class of "transparent potentials" that depend on α is obtained.

We noted that the method works for *almost* every set of phase shifts. To obtain the a_ℓ , we need $(1 + R)^{-1}$; on the other hand, once obtained, the c_ℓ yield the phase shifts δ_ℓ if the system of Eq. (32) has *only one* solution $A_\ell e^{i\delta_\ell}$. In both cases, the required conditions imply the regularity of matrices of the form $1 + MX$, where X is a diagonal matrix (with element equal to ic_ℓ , or to $\pm \mu_\ell \tan \delta_\ell$).

With the inequality in Eq. (45), the elements of X are bounded and MX is a completely continuous operator. Therefore, $1 + MX$ is invertible unless the phase shifts belong to *exceptional sets* for which MX has -1 as an eigenvalue.

An interesting consequence of the bounds required to make the method consistent is that the class of potentials that can be obtained is relatively narrow. For the convergence of Eq. (32), it is necessary that $|c_\ell| < C\ell$. It follows that $f_0^V(r, r')$ is an entire function of r and r' , that Eq. (24) has a unique solution for almost every r , and that $K(r, r')$ is a meromorphic function [Sabatier, 1966c]. Therefore, $V(r)$ is a meromorphic function, with poles of order 2, all bounded away from $r = 0$. Another limitation on the class of potentials comes from the fact that ℓ in Eq. (28) can be an integer only. Even potentials cannot be obtained in this way. For these

reasons, *Sabatier* [1967a] gave a generalization of the method by allowing ℓ in Eq. (28) to take any real values larger than $-1/2$. When this set S of values contain the integers and all the c_ℓ are absolutely bounded, the inverse problem can be solved and yields a much larger class of equivalent potentials. It can also be solved, at least when all the $\tan \delta_\ell$ are small enough, if ℓ is allowed to take half-integer values only [*Sabatier*, 1967a].

The bounds of the c_ℓ also yield interesting representations of the Jost function in the ℓ complex plane. It is convenient to use the notation μ for $\ell + 1/2$, $\gamma_\mu(V, 0)$ for $c_{\mu-1/2}$, γ_μ for $u_{\mu-1/2}$, ϕ_μ for $\psi_{\mu-1/2}$, so that $f_0^V(r, r')$ reads

$$f_0^V(r, r') = \sum_{\mu \in S_0} \gamma_\mu(V, 0) \gamma_\mu(r) \gamma_\mu(r') \quad (47)$$

Let us assume that the γ_μ are absolutely equibounded. The Jost functions are then given by [*Sabatier*, 1966c; 1967a]

$$f_{1,2}(\lambda) = \exp \left[\pm \frac{1}{2} i\pi \lambda - \frac{1}{2} \right] - \sum_{\mu \in S_0} \frac{\sin [(\lambda - \mu)(\pi/2)]}{\lambda^2 - \mu^2} \gamma_\mu(V, 0) f_{1,2}(\mu) \quad (48)$$

which can be interpreted as a special dispersion formula [*Newton*, 1967]. When S_0 contains all the numbers of the form n/k (k , fixed integer, $n = 1, 2, \dots$), it can be interpreted as a Lagrange-Valiron interpolation formula [*Sabatier*, 1967b] or a derived form of such an interpolation. Other interpolation formulas can be obtained for the wave functions which are valid even for γ_μ strongly increasing as $\mu \rightarrow \infty$. This interpretation led us to the additive formula

$$\gamma_\mu(W, V) = \gamma_\mu(W, V_0) - \gamma_\mu(V, V_0) \quad (49)$$

The interpolation formulas for the wave functions make it possible to derive the integral Eq. (24). Therefore, as pointed out and illustrated by *Sabatier* [1968], generalizations of the interpolation formulas may yield generalizations of the Regge-Newton procedure.

The Newton-Sabatier method can be used to construct exactly solvable potentials [*Newton*, 1962; *Sabatier*, 1967a]. In some cases, it is even possible to obtain a closed form of the scattering amplitude [*Sabatier*, 1966c; *Cox and Thompson*, 1968]. In addition, throughout this method, the first moment

$$\int_0^\infty rV(r) dr$$

can be related to the asymptotic behavior of the c_ℓ [*Sabatier*, 1967a]; in particular, if they vanish for $\ell > L < \infty$, this moment is zero [*Newton*, 1967]. Some authors have emphasized this point [*Cox and Thompson*, 1970a; *Coudray and Coz*, 1971] and have given a method [*Cox and Thompson*, 1970a,b] to avoid what they consider an essential complication for numerical calculations. In our opinion, this point is certainly over-emphasized (see below for numerical calculations). Its main effect is to produce oscillations in the tail of the potential.

Generalizations of the Regge-Newton Procedure

Relatively direct generalizations of the Regge-Newton procedure, in the framework of generalized translation operators, include those to Coulomb potentials [Coudray and Coz, 1970] and relativistic cases [Coudray and Coz, 1971]. Generalizations to spin-orbit potentials [Sabatier, 1968] and tensor forces [Hooshyear, 1970] necessitate much deeper transformations of the machinery. All these studies are still in a "formal" state, and much work is needed to clarify their details.

COMPLETE SOLUTION OF THE INVERSE PROBLEM AT FIXED ENERGY

The methods discussed in the preceding section yield answers only to questions (1), (2), and (3). In the "Regge-Loeffel" method, the answers to (1) and (2) are very general, but there is almost no answer to (3). In the methods using the coefficients c_μ , the answer to (3) is confined to special classes of potentials, which we generally denote as "the class C." In these methods, if we allow μ to have complex values, taking care of the reality of $V(r)$, it is certainly possible to obtain expansions of Eq. (27) that are complete in a large space of functions. At the same time, however, we preclude the use of the inverse-matrix procedures; furthermore, inclusion of the integral values of μ would give redundant expansions. *Sabatier* [1970, 1971] has proposed the following method of solution to the inverse problem, which allow answers to all six questions (1) through (6).

Let \mathcal{L} be the class of functions that are absolutely bounded by $Cr^{-1+\epsilon}$ and $Cr^{-3-\epsilon}$, and \mathcal{E} the class of potentials $V(r)$ such that $V(r)$, $rV'(r)$, and $r^2V''(r)$ belong to \mathcal{L} . It has been proved that for any potential of \mathcal{E} , the transformation kernel $K(r, r')$ and its derivative with respect to r can be written [Sabatier, 1970]

$$\left(1 - i \frac{\partial}{\partial r}\right) K(r, r') = -ir' \mathcal{K}(r') \exp[ir] + K_N(r, r') \quad (50)$$

where $K_N(r, r')$ is a so-called "negligible" function with respect to r , namely:

$$\left. \begin{aligned} K_N(r, r') &\rightarrow 0 & r \rightarrow \infty \\ \int_0^r |K_N(r, r')| [r'(1+r')]^{-1} dr' &\rightarrow 0 & r \rightarrow \infty \end{aligned} \right\} \quad (51)$$

Using Eqs. (21), (50), and (51), it is possible to obtain the remarkable formula

$$\int_0^\infty \mathcal{K}(\rho) u_\rho(\rho) \rho^{-1} d\rho = \exp\left(-i\ell \frac{\pi}{2}\right) [1 - A_\ell \exp(i\delta_\ell)] \quad (52)$$

The scattering structure function $\mathcal{K}(\rho)$ can be related bijectively to $V(r)$. The important step in the inverse problem therefore resides in obtaining $\mathcal{K}(\rho)$ from the δ_ℓ through Eq. (52). This is a kind of generalized moment problem and could be treated in this way [Sabatier, in preparation]. It is also possible to use an asymptotic form of a Gelfand-Levitan formulas [Sabatier, 1971]. For potentials of \mathcal{E} , the function $f_0 V(r, r')$ and its derivative can be expanded:

$$\left(1 - i \frac{\partial}{\partial r}\right) f_0^V(r, r') = -ir' \mathcal{F}(r') \exp(ir) + f_N(r, r') \quad (53)$$

where $f_N(r, r')$ is negligible with respect to r . From Eqs. (24), (50), and (53), we obtain

$$\mathcal{X}(r') = \mathcal{F}(r') - \int_0^\infty \mathcal{X}(\rho) \mathcal{F}(\rho, r') d\rho \quad (54)$$

where

$$\mathcal{F}(\rho, r') = (rr')^{-1} j(r, r') \quad (55)$$

Equation (54) supersedes the Regge-Newton Eq. (24) in all situations where r is largest — that is, in scattering studies. Again, for potentials of \mathcal{E} , it is possible to prove that [Sabatier, 1970, 1971]

$$\mathcal{F}(r, r') = 4 \text{ l.i.m. } \int_0^\infty w^{-1} \sin w u F(u) du \quad (56)$$

where

$$w = [(r - r')^2 + 4rr'u^2]^{1/2} \quad (57)$$

$$F(u) = -(2\pi)^{-1} V_0(u^{-1} - 1) \mathcal{O}(1 - u) + \psi(u) \quad (58)$$

V_0 is the first moment of the potential, \mathcal{O} the Heaviside function, $\psi(u)$ a function of $L_2(\mathcal{O}^\infty)$, and l.i.m. means limit in $L_2(\mathcal{O}^\infty)$. The premise of our method is that $F(u)$, which is the sine Fourier transform of $r\mathcal{F}(r, r)$, can be split into two parts:

$$F(u) = F(u) \mathcal{O}(1 - u) + F(u) \mathcal{O}(u - 1) \quad (59)$$

When inserted in Eq. (56), the two parts of the right-hand side of Eq. (59) give rise to two functions: $\mathcal{F}^I(r, r')$, whose spectrum is confined to $(0, 1)$; and $\mathcal{F}^E(r, r')$, whose spectrum is confined to $(1, \infty)$. These two functions are called the internal and external projections of $\mathcal{F}(r, r')$, respectively, and both can be expanded as in Eq. (53) to yield the internal and the external projections of $\mathcal{F}(r')$, say, $\mathcal{F}^I(r')$ and $\mathcal{F}^E(r')$. For convenience, we also define the slightly different functions

$$\begin{aligned} \overline{\mathcal{F}}^I(r, r') &= \mathcal{F}^I(r, r') + \delta (rr')^{-1} \sin r \sin r' \\ \overline{\mathcal{F}}^E(r, r') &= \mathcal{F}^E(r, r') - \delta (rr')^{-1} \sin r \sin r' \end{aligned} \quad (56)$$

and the corresponding quantities $\mathcal{F}^I(r')$ and $\mathcal{F}^E(r')$. We are now in position to solve Eq. (54). For this purpose we define $\mathcal{H}^I(r)$ and $\mathcal{H}^E(r)$ by

$$\mathcal{H}^I(r) = \mathcal{F}^I(r) - \int_0^\infty \mathcal{H}(\rho) \mathcal{F}^I(\rho, r) d\rho \quad (57)$$

$$\mathcal{H}^E(r) = \mathcal{F}^E(r) - \int_0^\infty \mathcal{H}(\rho) \mathcal{F}^E(\rho, r) d\rho \quad (58)$$

Clearly, Eq. (54) is satisfied if and only if

$$\mathcal{H}^E(r) + \mathcal{H}^I(r) = \mathcal{H}(r) \quad (59)$$

We now choose $\mathcal{F}^E(r, r')$ arbitrarily in a class of functions \mathcal{F} to which the functions $\mathcal{F}^E(r, r')$ corresponding to potentials of \mathcal{E} do belong. This class of functions is thoroughly studied in Sabatier [1971]. Let $\mathcal{G}^E(r, r')$ be the resolvent:

$$\mathcal{G}^E(r, r') = \mathcal{F}^E(r, r') - \int_0^\infty \mathcal{G}^E(r, \rho) \mathcal{F}^E(\rho, r') d\rho \quad (60)$$

Using Eq. (60), we can write Eq. (58) as

$$\mathcal{H}^E(r) = \mathcal{H}^E(r) - \int_0^\infty \mathcal{H}^I(\rho) \mathcal{G}^E(\rho, r) d\rho \quad (61)$$

where $\mathcal{H}^E(r)$ can be derived in a straightforward manner from $\mathcal{F}^E(r, r')$ and $\mathcal{G}^E(r, r')$, and therefore can be considered a known quantity as soon as $\mathcal{F}^E(r, r')$ has been chosen in \mathcal{F} .

By inserting Eqs. (61) and (59) into Eq. (57) we now can obtain an equation *reducible to an infinite system of linear algebraic equations that can be exactly solved!* Since

$$\mathcal{F}^I(r, r') = 4 \int_0^1 w^{-1} \sin w F(u) u du \quad (62)$$

we can write

$$\mathcal{F}^I(r, r') = (\pi')^{-1} \sum_0^{\infty} c_q u_q(r) u_q(r') \quad (63)$$

where

$$c_{\ell} = 4(2\ell + 1) \int_0^1 F(u) P_{\ell}(1 - 2u^2) u \, du \quad (64)$$

The coefficients c_{ℓ} are absolutely equibounded and go to a constant as $\ell \rightarrow \infty$.

Similar formulas can be obtained for $\mathcal{F}^I(r, r')$, and $\mathcal{F}^I(r)$. Inserting them in Eq. (57) and taking into account the information contained in Eq. (52'), we obtain the remarkable formula

$$\mathcal{H}^I(r) = \sum_{\ell=0}^{\infty} c_{\ell} A_{\ell} \exp \left[i \left(\delta_{\ell} - \ell \frac{\pi}{2} \right) \right] r^{-1} u_{\ell}(r) \quad (65)$$

which is independent of the choice of $\mathcal{F}^E(r, r')$.

Now inserting Eqs. (65), (59), and (61) into Eq. (52), we obtain the infinite system of equations

$$\begin{aligned} & A_{\ell} \exp \left[i \left(\delta_{\ell} - \ell \frac{\pi}{2} \right) \right] + B_{\ell} \exp \left[i \left(\sigma_{\ell} - \ell \frac{\pi}{2} \right) \right] \\ &= \exp \left(-i \ell \frac{\pi}{2} \right) - \sum_{\ell'} c_{\ell'} L_{\ell \ell'} A_{\ell'} \exp \left[i \left(\delta_{\ell'} - \ell' \frac{\pi}{2} \right) \right] + \sum_{\ell'} c_{\ell'} g_{\ell \ell'} A_{\ell'} \exp \left[i \left(\delta_{\ell'} - \ell' \frac{\pi}{2} \right) \right] \end{aligned} \quad (66)$$

where we set

$$B_{\ell} \exp \left[i \left(\sigma_{\ell} - \ell \frac{\pi}{2} \right) \right] = \int_0^{\infty} \mathcal{H}^E(\rho) u_{\ell}(\rho) \rho^{-1} \, d\rho \quad (67)$$

$$g_{\ell \ell'} = \int_0^{\infty} \int_0^{\infty} q^E(x, y) u_{\ell}(x) u_{\ell'}(y) x^{-1} y^{-1} \, dx \, dy \quad (68)$$

The quantities B_{ℓ} , σ_{ℓ} , $g_{\ell \ell'}$, are known after the choice of $\mathcal{F}^E(r, r')$. (Actually $g_{\ell \ell'}$ can be given an explicit form in terms of B_{ℓ} and σ_{ℓ}). If η_{ℓ} stands for:

$$\eta_{\ell} = B_{\ell} (\sin \sigma_{\ell} - \cos \sigma_{\ell} \tan \delta_{\ell}) \quad (69)$$

It is possible to show that η_{ℓ} vanishes more rapidly than ℓ^{-1} as $\ell \rightarrow \infty$, and to reduce the system to the form

$$(\eta + \tan \Delta) \mathbf{e} = M(1 + S) \mathbf{a} \quad (70)$$

where $a_\ell = c_\ell A_\ell \cos \delta_\ell$, and S is a conveniently defined matrix. As in the Newton-Sabatier method, it is possible to show that $1 + S$ is invertible for almost every set of phase shifts. Since S depends on the arbitrary function $\mathcal{F}^E(r, r')$ it is possible to choose it in such a way that $(1 + S)^{-1}$ exist and that the c_ℓ and the function $\mathcal{F}^E(r, r')$ again give the δ_ℓ .

Therefore, the method just described gives an answer to question (1) in a class of potentials that contains \mathcal{E} . It gives a negative answer to (2), and the solution depends on an arbitrary function of \mathcal{F} . It gives a positive answer to (3) in a class of potentials that contains \mathcal{E} . Since \mathcal{E} is hardly smaller than the class \mathcal{O} of potentials in which practically all the results of potential scattering have been obtained, these results are already useful. Moreover, the method gives a plain answer to (4) in the class \mathcal{E} . Still more interesting, it shows that the deviation of all potentials from each other can be analyzed in terms of the values of $F(u)$ between $u = 1$ and $u = \infty$. If we choose inside \mathcal{E} a class of potentials characterized, for instance, by given bounds on the derivatives, the importance of $F(u)$ between 1 and ∞ is smaller for smaller bounds, and in any case decreases and goes to zero as $E \rightarrow \infty$. Therefore, the deviation of equivalent potentials from each other is smaller when they are more "smooth," and in any case becomes smaller and smaller as E increases. These results have been sketched and checked through numerical computations by *Sabatier and Quyen-Van-Phu* [1970]. Except for a preliminary attempt by *Sabatier* [1967c], the last reference is the only available study on the deviation of equivalent potentials from each other. It is also the only available answer to question (5), and it paves the way for practical uses of the inversion procedure.

Detailed answers to question (6) are still lacking. A general answer can however be given using the method above. Since this method is constructive, it is easy to check the continuity of all of its steps, in convenient normed spaces and hence to give, in general, an affirmative answer to question (6), once $F(u)$ has been chosen for $u \in (1, \infty)$.

DISCUSSION

Calogero: Within a certain class of potentials, Martin is able to solve all the problems you have mentioned. He is able to prove existence, prove uniqueness, etc.

Moses: You don't use analytic continuation in ℓ . Have you tried using this in the one-dimensional problem? This has application to electromagnetic theory.

Sabatier: There is a difference between this problem and the one at fixed ℓ .

EXPLICIT EXPRESSIONS OF THE POTENTIAL AND ITS DERIVATIVES AT THE ORIGIN IN TERMS OF THE SCATTERING DATA

F. Calogero

Istituto di Fisica dell' Università di Roma

Istituto Nazionale di Fisica Nucleare, Sezione di Roma. Rome, Italy

ABSTRACT

N73-11610

The framework of this review paper is the quantum mechanical theory of scattering of a particle by a spherically symmetrical potential. As in the inverse scattering problem, the input of the calculation is the scattering and bound-state data, and the output is data on the potential. The results discussed are explicit expressions for the values of the potential and its derivatives at the origin in terms of the scattering and bound-state data. Various methods to obtain these results are outlined. The presentation is aimed at introducing these various approaches; the simplest scattering problem (nonrelativistic *S*-wave scattering on a holomorphic potential without bound states) is used as the basis for discussion, and technicalities are omitted whenever possible without loss of clarity. A complete compilation is given of the results obtained to date in this field, including the treatment of higher partial waves and the Klein-Gordon and Dirac equations.

INTRODUCTION

The study of the inverse problem in scattering theory consists in the development of mathematical techniques to obtain information on the cause of the scattering from the parameters that are measured in a scattering experiment. In the framework of quantum mechanics, major attention has been given to the case of a non-relativistic particle interacting, via a potential, with a scattering center; or, equivalently, that of two particles interacting through a potential depending on their relative distance, a problem that reduces to the previous one after the (trivial) center-of-mass motion has been separated out. Even for this specific case, several different "inverse problems" can be stated, depending on the input information taken as the starting point of the calculation. This paper focuses on the problem in which the input information is one scattering phase shift (assumed known for all energies) and the parameters of the corresponding bound states.

This inverse scattering problem has been solved by I. M. Gelfand and B. M. Levitan and by V. A. Marchenko; a detailed account of it is given by Newton in the opening paper of this chapter. The assertion that this problem has been solved means that existence and uniqueness properties have been stated, and that procedures for the reconstruction of the potential function from the input data have been given. These procedures involve quadratures and the solution of an integral Fredholm-type equation.

The developments reviewed in this paper constitute an attempt to bypass the practical difficulty implied by the need to solve an integral equation; the goal is to obtain information on the potential in the form of *explicit* expressions involving the scattering and bound-state data. Specifically, attention concentrates on the values of the potential and of its derivatives at the origin, and explicit closed expressions for these quantities are obtained in terms of integrals and sums over the scattering and bound-state data.

To obtain these results, different routes may be followed. In keeping with the multidisciplinary character of this chapter, the emphasis here is on the illustration of various possibilities that might be suggestive of

developments in other fields, rather than on the presentation of the most general and rigorous picture. We therefore consider the simplest possible problem (*S*-wave scattering on a holomorphic potential without bound states), omitting technical details whenever this is possible without loss of clarity. We also review all the results that this line of research has yielded up to now, referring for the corresponding proofs and more detailed presentation to the original papers.

The line of research covered in this review originates from a result obtained simultaneously and independently by *Newton* [1956] and *Faddeev* [1957] — a closed expression for the value of the potential at the origin as a sum over the *ℓ*-wave bound-state energies plus an integral over the *ℓ*-wave scattering phase shift, considered as a function of energy. In a sense, *Levinson's theorem* [*Levinson*, 1949] was already a first instance of this kind of result. Subsequent developments are due to *Buslaev and Faddeev* [1960], *Buslaev* [1962], *Percival* [1962], *Percival and Roberts* [1963], *Roberts* [1963; 1964b]. These authors, employing different approaches, obtained results that implied the possibility of obtaining all the derivatives of the potential at the origin in terms of one scattering phase shift and the corresponding bound-state energies, provided the potential is an even function of *r* (so that all its derivatives of odd order vanish). However the conditions to be satisfied by phase shifts for the potential function to be even could not be obtained. The solution of the problem for general potentials and *S* waves was given by *Calogero and Degasperis* [1968]; this work also yields the conditions that the *S*-wave phase shift must satisfy for the corresponding potential to be even. The extension of these results to the cases of the Klein-Gordon and Dirac equations are due to *Degasperis* [1970] and to *Corbella* [1970], respectively, and the treatment of higher partial waves is due to *Corbella* [1971].

The techniques discussed here produce interesting results other than those specifically cited and can be applied in a more general framework than that considered in this paper. *Buslaev* [1966], for example, treats the case of scattering in the three-dimensional context, without performing a partial-wave expansion, and elucidates the explicit relationships between the scattering amplitude and the potential that are implied by the trace method. *Buslaev's* paper provides a complete review up to 1966, including the numerous important contributions (not mentioned here) of the Russian school of mathematical physics and functional analysis; it is written in accord with standards of mathematical rigor that are considerably more stringent than those adopted in the present paper.

POTENTIAL SCATTERING

The stationary Schrödinger equation in three dimensional space reads

$$[-\Delta + V(r)] \psi(r) = E \psi(r) \quad (1)$$

with units chosen such that $\hbar^2/(2m) = 1$. We assume hereafter that the potential $V(r)$ is spherically symmetrical, that it vanishes at infinity faster than r^{-2} ,

$$\lim_{r \rightarrow \infty} [r^2 V(r)] = 0 \quad (2)$$

that it is finite valued (for nonnegative values of *r*), and that it is an entire function of *r*, so that the expansion

$$V(r) = \sum_{n=0}^{\infty} \frac{V_n r^n}{n!} \quad (3)$$

is convergent for all values of r . The quantity v_n is clearly the n th derivative of the potential, evaluated at the origin. This last assumption is certainly much more stringent than is required for the validity of all the results given below. Presumably it would be sufficient to assume that $V(r)$ and all its derivatives are finite-valued for real nonnegative r .

The scattering solution of the stationary Schrödinger equation is characterized by the asymptotic boundary condition

$$\psi_k(r) \xrightarrow{r \rightarrow \infty} \text{const} \exp(i \vec{k} \cdot \vec{r}) + f(k, \theta) \frac{\exp(ikr)}{r} \quad (4)$$

with

$$k^2 = E \quad (5)$$

and

$$\cos \theta = \frac{\vec{k} \cdot \vec{r}}{kr} \quad (6)$$

Thus this solution contains asymptotically a plane wave $\exp(i \vec{k} \cdot \vec{r})$, corresponding to the incoming beam, and an outgoing scattered wave, whose angular distribution is characterized by the "scattering amplitude" $f(k, \theta)$. The differential (elastic) scattering cross section $\sigma(E, \theta)$ – the measurable quantity – is related to the scattering amplitude by

$$\sigma(E, \theta) = |f(k, \theta)|^2 \quad (7)$$

The cylindrical symmetry of the problem suggests the introduction of the "radial wave functions" $\phi_l(k, r)$ through the position

$$\psi_k(r) = \sum_{l=0}^{\infty} \frac{c_l P_l(\cos \theta) \phi_l(k, r)}{r} \quad (8)$$

where the P_l are Legendre polynomials. The Schrödinger Eq. (1) then yields the "radial Schrödinger equation"

$$\phi_l''(k, r) + [k^2 - l(l+1)r^{-2} - V(r)] \phi_l(k, r) = 0 \quad (9)$$

with the boundary condition

$$\phi_l(k, 0) = 0 \quad (10)$$

In Eq. (9) and the following, primes indicate differentiation with respect to the last argument. The hypothesis on the potential introduced above ensures that this boundary condition characterizes uniquely, up to a multiplicative constant, a solution of Eq. (9). They also imply for the asymptotic behavior of the radial wave functions $\phi_\ell(k, r)$ the form

$$\phi_\ell(k, r) \xrightarrow{r \rightarrow \infty} \text{const} \times \sin \left[kr - \ell \frac{\pi}{2} + \eta_\ell(k) \right] \quad (11)$$

The quantities $\eta_\ell(k)$, called *scattering phase shifts*, are related to the scattering amplitude by the equation

$$f(k, \theta) = k^{-1} \sum_{\ell=0}^{\infty} (2\ell + 1) \exp[i\eta_\ell(k)] \sin \eta_\ell(k) P_\ell(\cos \theta) \quad (12)$$

which follows from Eq. (11), (8), and (4). Thus, knowledge of all the scattering phase shifts is tantamount to knowledge of the scattering amplitude of the scattering amplitude if, and vice versa, since obviously

$$\exp[i\eta_\ell(k)] \sin \eta_\ell(k) = \frac{k}{2} \int_0^\pi d\theta \sin \theta P_\ell(\cos \theta) f(k, \theta) \quad (13)$$

Note that in the sum of Eq. (12) only the terms with

$$\ell \leq L \simeq k\bar{r} \quad (14)$$

where \bar{r} is a measure of the range of the potential, contribute appreciably; in fact for $\ell \gg L$ the phase shifts $\eta_\ell(k)$ become negligibly small. Thus, at low energy only the phase shift with $\ell = 0$ — the *S-wave* phase shift — is important, and of course the scattering becomes isotropic.

If the potential has a sufficiently strong attractive (negative) part, the radial Schrödinger equations (9) may have normalizable solutions for some discrete nonpositive values of the energy $E = k^2$. The conditions on the potential that we have assumed are sufficient to ensure that at most a finite number of such solutions can exist. They correspond to "bound states." Every bound state is characterized by its (negative) energy $E_{\ell n}$, and by the "normalization constant" $C_{\ell n}$ defined by

$$C_{\ell n} = \left[\int_0^\infty dr \phi_\ell^2(E_{\ell n}, r) \right]^{-1} \quad (15)$$

where $\phi_\ell(E_{\ell n}, r)$ is the bound-state wave function, normalized so that

$$\lim_{r \rightarrow 0} \left[\frac{\phi_\ell(E_{\ell n}, r)}{r^{\ell+1}} \right] = [(2\ell + 1)!!]^{-1} \quad (16)$$

A sufficient condition to ensure that no bound states be present is the inequality

$$-\int_0^{\infty} dr r V(r) \theta[-V(r)] < 1 \quad (17)$$

where $\theta(x)$ is the usual step function,

$$\begin{aligned} \theta(x) &= 1 & x > 0 \\ \theta(x) &= 0 & x < 0 \end{aligned} \quad (18)$$

The inverse scattering problem consists in the determination of the potential $V(r)$ from the phase shifts $\eta_l(k)$ (considered as a known function of k), and from the energies E_{ln} and the normalization constants C_{ln} , for a particular value of l . These parameters are, in fact, sufficient to determine uniquely the potential, as described in detail by Newton at the beginning of this chapter.

In the rest of this section, and in the following three, we consider the case of S waves ($l = 0$), omitting systematically the subscript zero. We also assume, for simplicity, that no bound states are present (except in the discussion of the method of operator traces).

In addition to the radial Schrödinger wave function characterized by the boundary conditions of Eq. (10), it is expedient to introduce the Jost solution $f(k, r)$ of the radial Schrödinger Eq. (9) (with $l = 0$), characterized by the asymptotic boundary condition

$$\lim_{r \rightarrow \infty} [f(k, r) \exp(ikr)] = 1 \quad (19)$$

Its value at the origin is the Jost function

$$f(k) = f(k, 0) \quad (20)$$

The phase of this function is just the scattering phase shift

$$f(k) = |f(k)| \exp[i\eta(k)] \quad (21)$$

Its modulus can be expressed through the phase shift by

$$|f(k)| = \exp[\Delta(k)] \quad (22)$$

where

$$\Delta(k) = -\frac{2}{\pi} P \int_0^{\infty} dq \frac{q}{q^2 - k^2} \eta(q) \quad (23)$$

and P indicates the principal value of the integral. Of course, Eqs. (21) through (23) are written for real k . The inverse square of the modulus of the Jost function, or more precisely the quantity $(1 - |f(k)|^{-2})$, is termed the *spectral function*, and it will play an important role in the following. Another important auxiliary function is

$$g(k) = ik + \frac{d}{dr} \ln f(k, r) \Big|_{r=0} \quad (24)$$

For real k , the imaginary part of this function is simply related to the spectral function by

$$\text{Im } g(k) = k [1 - |f(k)|^{-2}] \quad (25)$$

which follows from the definition of the Jost function through the boundary condition in Eq. (19), implying for real k and r

$$f(-k, r) = [f(k, r)]^* \quad (26)$$

and from the Wronskian relation

$$f'(k, r)f(-k, r) - f(k, r)f'(-k, r) = 2ik \quad (27)$$

The definition of $f(k, r)$ moreover implies that this is a holomorphic function of k in the half plane $\text{Im } k \leq 0$, and that the only zeros of $f(k)$ that might occur in this half plane are located on the imaginary axis [as is easily proved from the differential Eq. (9) and the boundary condition in Eq. (19)] and correspond to bound states. Indeed if

$$f(k, r) \Big|_{r=0} = f(k) = 0 \quad \text{for} \quad k = -ip, \quad p > 0 \quad \text{then} \quad f(k, r)$$

satisfies the boundary condition of Eq. (10) characterizing the radial wave function, and at the same time the boundary condition in Eq. (19) implies that it vanishes asymptotically proportionally to $\exp(-pr)$; thus, $f(-ip, r)$ coincides, up to a multiplicative constant, with the normalizable wave function $\phi(E, r)$ of a bound state, with $E = -p^2$. Thus if no bound states are present, $g(k)$ is also holomorphic in the lower half of the complex k plane, and this (together with its asymptotic properties discussed below) implies that for real k its real part can be expressed as an integral over its imaginary part, namely

$$\text{Re } g(k) = \frac{2}{\pi} P \int_0^\infty dq \left[\frac{q^2}{q^2 - k^2} \right] [|f(q)|^{-2} - 1] \quad (28)$$

To obtain this equation we have also used the reflection property

$$\text{Im } g(-k) = -\text{Im } g(k) \quad (29)$$

that follows from Eq. (25) and from

$$|f(-k)| = |f(k)| \quad (30)$$

This last equation is implied by Eqs. (26) and (21) together with

$$\eta(-k) = -\eta(k) \quad (31)$$

As for the reflection property of the real part of $g(k)$, clearly Eq. (28) implies

$$\operatorname{Re} g(-k) = \operatorname{Re} g(k) \quad (32)$$

Equations (28) through (32) are, of course, for real k .

The convergence of the integrals in Eqs. (23) and (28) is implied by the asymptotic vanishing of the phase shift $\eta(k)$ and of the spectral function $1 - |f(k)|^{-2}$. In fact, the properties that we have assumed for the potential imply that the phase shift $\eta(k)$ admits the asymptotic expansion [Verde, 1955]

$$\eta(k) = \sum_{n=0}^N a_n k^{-2n-1} + O(k^{-2N-3}) \quad (33)$$

and this result with Eqs. (22) and (23) implies for $\Delta(k)$ the asymptotic expansion

$$\Delta(k) = \sum_{n=1}^N d_n k^{-2n} + O(k^{-2N-2}) \quad (34)$$

and therefore for $|f(k)|$ the asymptotic behavior

$$|f(k)| = 1 + O(k^{-2}) \quad (35)$$

A more detailed discussion of these asymptotic expansions is given in a later section...

THE METHOD OF FUNCTIONAL INTEGRATION

On the basis of the Gelfand-Levitan solution of the inverse scattering problem, Newton [1956, 1966] obtained the following expression for the functional derivative of the potential $V(r)$ with respect to the scattering phase shift $\eta(k)$:

$$\frac{\delta V(r)}{\delta \eta(k)} = -\frac{4}{\pi} k \frac{d}{dr} G(k^2, r) \quad (36)$$

where

$$G(k^2, r) = -(2k)^{-1} \operatorname{Im} \left\{ \exp [2i\eta(k)] f^2(-k, r) \right\} \quad (37)$$

and $f(k, r)$ is the Jost solution of the radial Schrödinger equation. In writing these equations, and always in this section, we assume k to be real. The fact that $G(k^2, r)$ is a function of k^2 rather than k is implied by the symmetry and reality properties of $\eta(k)$ and $f(k, r)$ [Eqs. (26), (31), and (21)].

The validity of Eq. (36) is elucidated in the following, when we tackle the problem of its integration. For the moment we limit ourselves to specify that whenever we write a functional derivative, we understand the momentum variable k to extend over the range $-\infty, +\infty$; the values of the scattering phase shift for negative k are obtained through the reflection formula in Eq. (31) from the "experimental" values, corresponding by convention to non-negative k [see Eq. (4)] (note that for positive k the scattered wave is outgoing, because the time-dependence characteristic of a stationary function consists of the multiplicative factor $\exp(-ik^2 t)$).

We now note that the definition of $G(k^2, r)$ [Eq. (37)] and the differential Eq. (9) satisfied by $f(-k, r)$ imply that $G(k^2, r)$ satisfies the third-order equation

$$G'''(k^2, r) = 4[V(r) - k^2] G'(k^2, r) + 2V'(r)G(k^2, r) \quad (38)$$

with boundary conditions

$$G(k^2, 0) = 0 \quad (39)$$

$$G'(k^2, 0) = -1 \quad (40)$$

$$G''(k^2, 0) = -2\operatorname{Re} g(k) \quad (41)$$

We introduce now the power expansion of $G(k^2, r)$:

$$G(k^2, r) = \sum_{n=1} \frac{G_n(k^2) r^n}{n!} \quad (42)$$

The convergence of this expansion is implied by the assumed holomorphy of the potential [Verde, 1955]. Introducing this expansion, together with that of the potential [Eq. (3)], in the differential Eq. (38), we get the recursion relations

$$G_{n+3}(k^2) = -4k^2 G_{n+1}(k^2) + 2 \sum_{m=1}^{n+1} \frac{n+m+1}{n!} \binom{n+1}{m} G_m(k^2) V_{n+1-m} \quad (43)$$

while the boundary conditions in Eqs. (39) through (41) imply

$$G_1(k^2) = -1 \quad (44)$$

$$G_2(k^2) = -2\operatorname{Re} g(k) \quad (45)$$

But from Eq. (36) we infer that

$$\frac{\delta V_n}{\delta \eta(k)} = -\frac{4}{\pi} k G_{n+1}(k^2) \quad (46)$$

and inserting this relation in the recursion relations of Eq. (43) we obtain

$$\frac{\delta V_{n+2}}{\delta \eta(k)} = -4k^2 \frac{\delta V_m}{\delta \eta(k)} + 2 \sum_{m=0}^n \frac{n+m+2}{n+1} \binom{n+1}{m+1} V_{n-m} \frac{\delta V_m}{\delta \eta(k)} \quad (47)$$

Here, of course, the quantities V_n are considered functionals of the phase shift; recall that because we are limiting our consideration to the case without bound states, to every $\eta(k)$ there corresponds a unique potential $V(r)$, and therefore definite values for all the quantities V_n .

This system of recursive functional differential equations, together with the two starting conditions

$$\frac{\delta V_0}{\delta \eta(k)} = \frac{4}{\pi} k \quad (48)$$

and

$$\frac{\delta V_1}{\delta \eta(k)} = -\frac{8k}{\pi^2} P \int_{-\infty}^{+\infty} dq \frac{q^2}{q^2 - k^2} \left\{ 1 - \exp \left[\frac{2}{\pi} P \int_{-\infty}^{+\infty} dp \frac{\rho}{\rho^2 - q^2} \eta(p) \right] \right\} \quad (49)$$

implied by Eqs. (46), (44), (45), (28), (22) and (23), allow, by sequential functional integration, evaluation of the quantities V_n —namely, the values of the potential and its derivatives at the origin, in terms of the scattering phase shift $\eta(k)$. The functional integration is accomplished through the formula [Volterra, 1959]

$$V_n = \bar{V}_n + (s - \bar{s})^{-1} \int_{\bar{s}}^s dt \int_{-\infty}^{+\infty} dk [\eta(k) - \bar{\eta}(k)] \frac{\delta V_n}{\delta \eta(k)} \bigg|_{\eta'(k) = (s - \bar{s})^{-1} [(t - \bar{s})\eta(k) - (t - s)\bar{\eta}(k)]} \quad (50)$$

where V_n is the n th derivative (evaluated at the origin) of the potential $\bar{V}(r)$ corresponding to the phase shift $\bar{\eta}(k)$, just as V_n is the n th derivative of the potential $V(r)$ corresponding to the phase shift $\eta(k)$.

A difficulty is apparent, however, from the structure of this equation. Clearly, to be valid, the asymptotic vanishing at large k of the difference $\eta(k) - \bar{\eta}(k)$ must be sufficiently rapid. Using the fact that

$$f(k, r) = e^{-ikr} [1 + O(k^{-1})] \quad (51)$$

at large k , we get from Eqs. (36) and (37) the estimate, valid at large k ,

$$\frac{\delta V(r)}{\delta \eta(k)} = \frac{4}{\pi} k \cos(2kr) \quad (52)$$

and from this we get by differentiation

$$\frac{\delta V^{(n)}(r)}{\delta \eta(k)} = (-)^{n/2} \frac{4}{\pi} 2^n k^{n+1} \cos(2kr) \quad \text{for even } n \quad (53)$$

$$\frac{\delta V^{(n)}(r)}{\delta \eta(k)} = (-)^{(n+1)/2} \frac{4}{\pi} 2^n k^{n+1} \sin(2kr) \quad \text{for odd } n \quad (54)$$

where $V^{(n)}(r)$ indicates of course the n th derivative of $V(r)$. We therefore conclude that a sufficient condition for the validity of Eq. (50) is that

$$\eta(k) - \bar{\eta}(k) = O(k^{-n-3}) \quad (55)$$

at large k . Using the asymptotic expansion in Eq. (33) we see that this happens provided

$$a_j = \bar{a}_j, \quad j = 0, 1, \dots, n/2 \quad \text{for even } n \quad (56)$$

$$a_j = \bar{a}_j, \quad j = 0, 1, \dots, (n+1)/2 \quad \text{for odd } n \quad (57)$$

We are finally in the position to begin the process of functional integration. Applying the formula (50) to Eq. (48) we get

$$V_0 = \bar{V}_0 + \frac{4}{\pi} \int_{-\infty}^{+\infty} dk k [\eta(k) - \bar{\eta}(k)] \quad \text{provided } \bar{a}_0 = a_0 \quad (58)$$

or, equivalently,

$$V_0 - \frac{4}{\pi} \int_{-\infty}^{+\infty} dk k \left[\eta(k) - \frac{a_0}{k} \right] = \bar{V}_0 - \frac{4}{\pi} \int_{-\infty}^{+\infty} dk k \left[\bar{\eta}(k) - \frac{\bar{a}_0}{k} \right] \text{ provided } a_0 = \bar{a}_0 \quad (59)$$

the constants a_0 , \bar{a}_0 being defined by Eq. (33) and by a similar formula for $\bar{\eta}(k)$ in which the constants \bar{a}_n appear. This equation may also be written in the form

$$V_0 - \frac{4}{\pi} \int_{-\infty}^{+\infty} dk k \left[\frac{\eta(k) - a_0}{k} \right] = A_0(a_0) \quad (60)$$

where the function A_0 is a universal function of its argument. This function may be easily evaluated using a simple potential whose phase shift is exactly known, and it turns out to vanish identically (when no bound states are present). Therefore, we can write

$$V_0 = \frac{4}{\pi} \int_{-\infty}^{+\infty} dk k \left[\frac{\eta(k) - a_0}{k} \right] \quad (61)$$

$$= \frac{4}{\pi} \int_{-\infty}^{+\infty} dk \left(-k \frac{d}{dk} \right) [k\eta(k)] \quad (62)$$

The equivalence of Eq. (62) to Eq. (61) is easily checked by partial integration (see Appendix A).

Equations (61) and (62) provide an explicit expression for the value of the potential at the origin. They were first obtained by *Newton* [1956], using the same procedure as here, and by *Faddeev* [1957]. Both authors obtained a more general version; for instance, Newton's result can be written:

$$V_0 = \frac{8}{\pi} (2\ell + 1)^{-1} \int_0^{\infty} dk \left(-k \frac{d}{dk} \right) [k\eta_\ell(k)] + 4(2\ell + 1)^{-1} \sum_n E_{\ell n} \quad (63)$$

where the sum is extended over the energies of all the bound states with given ℓ . This formula is more general than Eqs. (61) and (62) in two respects: it includes the possibility that bound states are present, and it applies to all values of ℓ , not only the S -wave case ($\ell = 0$).

Continuing the process of sequential integration, and applying the rule (50) to Eq. (49), one obtains

$$V_1 = \bar{V}_1 - \frac{4}{\pi} \int_{-\infty}^{+\infty} dk k^2 [|f(k)|^{-2} + 2\Delta(k) - |\bar{f}(k)|^{-2} - 2\bar{\Delta}(k)] \text{ provided } a_0 = \bar{a}_0, a = \bar{a}_1 \quad (64)$$

where $\Delta(k)$ and $|f(k)|$ are given in terms of $\eta(k)$ by Eq. (22) and (23), and $\bar{\Delta}(k)$, $|\bar{f}(k)|$ are given by similar equations in terms of $\bar{\eta}(k)$. Again one can conclude from this formula that

$$V_1 + \frac{4}{\pi} \int_{-\infty}^{+\infty} dk k^2 [|f(k)|^{-2} + 2\Delta(k) - 1] = A_1(a_0, a_1) \quad (65)$$

where $A_1(a_0, a_1)$ is a universal function of its arguments. This function could be found using a simple solvable potential (containing at least two parameters), but we defer to a later section a derivation of the final form of this equation. Note that the integral in Eq. (65) is convergent, because the integrand vanishes asymptotically as k^{-2} ; this is implied by Eq. (22) and (34).

The process of sequential integration can be continued, using the recursion relations (47) and the explicit expressions of the V_n that are successively determined. For instance, the next functional equation to be integrated would be

$$\frac{\delta V_2}{\delta \eta(k)} = -\frac{16}{\pi} \left\{ k^3 + \frac{4}{\pi} \int_{-\infty}^{+\infty} dq \, q \frac{d}{dq} [q\eta(q)] \right\} \quad (66)$$

Clearly the formula (50) is always adequate to perform the integration, although the labor involved gets progressively more difficult.

THE METHOD OF OPERATOR TRACES

A complete treatment of the method of operator traces has been given by *Buslaev* [1966] in the context of three-dimensional scattering theory. Here we try to convey the main idea of this method in the most elementary setting, following the work of *Percival* [1962] and *Buslaev and Faddeev* [1960]. In terms of the problem under consideration, the main limitation of this approach is its inability to yield an expression for the first derivative of the potential at the origin, and in general to deal with potentials that are not even (so that their derivatives of odd order do not all vanish at the origin). On the other hand the great mathematical generality of the approach has a great potential for applications in other fields of mathematical physics. *Dikii* [1958] has reviewed the trace method in the general context of the theory of Sturm-Liouville differential operators.

The basic idea of this approach is best introduced considering the hamiltonian H of a quantum-mechanical system that admits only a finite number N of linearly independent states. Such a hamiltonian may be represented by an $N \times N$ matrix, and the invariance of the trace of a matrix under unitary transformations implies the well-known formula

$$\text{tr} [H^p] = \sum_{n=1}^N E_n^p \quad (67)$$

where the E_n 's are the eigenvalues of H and p is an arbitrary (integral) exponent.

Our problem is characterized by the hamiltonian

$$H = T + V \quad (68)$$

$$= -\frac{d^2}{dr^2} + V(r) \quad (69)$$

whose eigenvalues are all the positive values $E = k^2$, and possibly a finite number of negative values E_n , corresponding to (*S*-wave) bound states, as well. Here we keep open the possibility that bound states exist, since their presence does not cause any additional complication – on the contrary, it may even make the treatment more transparent.

There is now a continuous infinity of eigenvalues, so that the formula equivalent to Eq. (67) reads

$$\text{tr} [H^p] = \sum_n (E_n)^p + \int_0^\infty dE E^p n(E) \quad (70)$$

where $n(E)$ is the density of states in the continuum. This formula, however, has only a symbolic meaning, because the integral in the right-hand side generally does not converge (nor, of course, does the trace in the left-hand side). On the other hand, because the density of states $n(E)$ is simply related to the scattering phase shift $\eta(k)$ (see below), if it were possible to assign a definite meaning to this formula, one would obtain relationships between the energy moments of the scattering phase shift on the right-hand side, and the potential that is contained in the hamiltonian operator on the left-hand side. This motivates the attempt to extract definite relationships from the symbolic Eq. (70).

The first trick to use is to subtract the same expression for the case without interaction from Eq. (70), obtaining

$$\text{tr} [H^p - T^p] = \sum_n (E_n)^p + \int_0^\infty dE E^p n_f(E) \quad (71)$$

where $n_f(E)$ is now the increment in the density of states due to the interaction. It is well known (see Appendix B) that

$$n_f(E) = (2\pi k)^{-1} \frac{d}{dk} \eta(k) \quad (72)$$

where $\eta(k)$ is the scattering phase shift. Thus, in place of Eq. (71) we may write

$$\text{tr} [H^p - T^p] = \sum_n (E_n)^p + \frac{1}{\pi} \int_0^\infty dk k^{2p} \frac{d}{dk} \eta(k) \quad (73)$$

Of course this equation has still only a symbolic meaning, because in general the integral in the right-hand side does not converge. For $p = 0$, however, the equation is directly meaningful, and it implies the result

$$\eta(0) - \eta(\infty) = N\pi \quad (74)$$

where N is the number of (*S*-wave) bound states. This is just Levinson's theorem [Levinson, 1949].

We now outline the two approaches of Percival [1962] and Buslaev and Faddeev [1960] to extract meaningful relationships from the symbolic Eq. (73).

Percival Approach

In place of Eq. (73) one substitutes the equation

$$\text{tr} \left[e^{-\beta H} - e^{-\beta T} \right] = \sum_n e^{-\beta E_n} + \frac{1}{\pi} \int_0^\infty dk e^{-\beta k^2} \frac{d}{dk} \eta(k) \quad (75)$$

which is now certainly meaningful for positive β . This approach is very close to that employed in certain problems of statistical mechanics; indeed, this equation can be found in an old paper by *Uhlenbeck and Beth* [1936], where it is used to study the virial expansion for a nonideal gas. It is then noted that, while an expansion of Eq. (75) in powers of β is not allowed, as it would merely reproduce Eq. (73), one can expand in powers of $\beta^{1/2}$ and equate the coefficients of equal powers of $\beta^{1/2}$. (This corresponds, in the context of statistical mechanics, to a rearrangement of the Kirkwood-Wigner expansion of the partition function [Landau and Lifshits, 1959]). In this manner we obtain the moment relationships

$$\begin{aligned} & -\frac{1}{4\pi} \int_{-\infty}^{+\infty} dr \int_{-\infty}^{+\infty} dk \exp(ikr) [(T+V)^p \cdots T^p] \exp(ikr) \\ & = \sum_n (E_n)^p + \frac{1}{\pi} \int_0^\infty dk k^{2p} \left[\frac{d\eta(k)}{dk} + \sum_{n=0}^{p-1} (2n+1) a_n k^{-2n-2} \right] \quad \text{for } p = 1, 2, 3, \dots \end{aligned} \quad (76)$$

The left-hand side originates from the trace operation; the right-hand side contains the coefficients a_n characterizing the asymptotic behavior of $\eta(k)$ through Eq. (33). Using the results of Appendices A and C, Eq. (76) can be rewritten in the more compact form

$$-\frac{1}{4} \left\{ 1 \cdot \left[V(r) - \frac{1}{4} (\vec{d} - \overleftarrow{d}) \cdot 1 \right]^p \right\} \Big|_{r=0} = \sum_n (E_n)^p + \left[\frac{2p}{\pi(2p-1)!} \right] \int_0^\infty dk D^{2p-1} \eta(k) \quad \text{for } p = 1, 2, 3, \dots \quad (77)$$

where the operators \vec{d} , \overleftarrow{d} and D are defined by

$$f(r) \vec{d} g(r) \equiv f(r) g'(r), f(r) \overleftarrow{d} g(r) \equiv f'(r) g(r), \quad \text{and} \quad D \equiv k \frac{d}{dk} k$$

(see Appendices A and C).

To obtain Eq. (76), and therefore Eq. (77), however, it is necessary to assume that the potential function $V(r)$ is not only holomorphic, but also even

$$V_{2n+1} \equiv \frac{d^{2n+1} V(r)}{dr^{2n+1}} \Big|_{r=0} = 0 \quad \text{for } n = 0, 1, 2, \dots \quad (78)$$

The values of $V(r)$ for negative r occurring in the integral in the left-hand side of Eq. (76) are then defined by

$$V(-r) = V(r) \quad (79)$$

This prescription, together with the asymptotic vanishing of $V(r)$, ensures convergence of the integration over the variable r .

The treatment of Appendix C implies that Eq. (77) can be rewritten in the form

$$V_{2p} = (-)^{p+1} \frac{2^{2p+2}}{p+1} \sum_n (E_n)^p + (-)^{p+1} \frac{2^{2p+3}}{\pi(2p+1)!} \int_0^\infty dk D^{2p+1} \eta(k) + F_p(V_0, V_2, \dots, V_{2p-2}) \quad (80)$$

where F_p is defined by the equation

$$F_p(V_0, V_2, \dots, V_{2p-2}) = (-)^{p+1} \frac{2^{2p}}{p+1} \left\{ 1 \cdot \left[V(r) - \frac{1}{4} (\vec{d} - \vec{d}')^2 \right]^{p+1} \cdot 1 \right\} \Big|_{r=0} + V_{2p} \quad (81)$$

and depends only on the V_n whose order n does not exceed $2p-2$. It is therefore clear that from Eq. (80) it is in principle possible to obtain sequentially all the even derivatives of the potential in terms of the generalized moments

$$L_p = \frac{2}{p} \sum_n (E_n)^p + \left[\frac{4}{\pi(2p-1)!} \right] \int_0^\infty dk D^{2p-1} \eta(k), \quad p = 1, 2, 3, \dots \quad (82)$$

The derivative V_{2p} of order $2p$ depends only on the moments whose order n does not exceed $p+1$. Of course, the knowledge of all the even derivatives is sufficient to reconstruct the whole potential since it is assumed to be even. On the other hand, the conditions that the phase shift and the bound-state parameters must satisfy so that the corresponding potential be even are far from trivial; moreover, their investigation is beyond the power of the approach we have just described.

Explicit expressions of the first derivatives of the potential, evaluated at the origin, in terms of the phase shift and the bound-state data, are given in a later section (for the general case of a potential that need not be even); therefore we do not report here the results that obtain from Eq. (80).

Buslaev and Faddeev Approach

The approach of *Buslaev and Faddeev* [1960], although originating from the method of operator traces, is, in fact, closer to the approach based on the asymptotic expansion of the scattering parameters, and on their analyticity properties (see the next section). This method is also applicable to potentials that are not even, but it can yield explicit expressions for all the derivatives of the potential at the origin only in the case of even potentials—that is, only if all the derivatives of odd order are a priori known to vanish at the origin.

The basic equation of the *Buslaev and Faddeev* [1960] method is

$$\sin(\pi z) \int_0^\infty dk k^{2z-1} \eta(k) + \cos(\pi z) \int_0^\infty dk k^{2z-1} \ln |f(k)| + \frac{\pi}{2z} \sum_n (-E_n)^z = 0 \quad (83)$$

that holds in the strip of the complex z plane characterized by the inequality

$$0 < \operatorname{Re} z < \frac{1}{2} \quad (84)$$

This equation is established by contour integration of the function $k^{2z}(d/dk) \ln f(k)$, exploiting the holomorphy of $f(k)$ in the lower half plane $\operatorname{Im} k < 0$, and the fact that its zeros there correspond to the bound states as indicated earlier.

The similarity of this equation to Eq. (73) should be noted, as also the disappearance of any explicit dependence on the interaction. Of course the restriction to the strip $0 < \operatorname{Re} z < 1/2$ is essential. The moment relationships are obtained by analytic continuation outside this strip, to the values $z = p + 1/2$ and $z = p$, with $p = 1, 2, 3, \dots$. This continuation leads to formulas that, using the results of Appendix A may be cast into the form

$$a_p = (-)^p \frac{2}{2p+1} \sum_n (-E_n)^{p+1/2} - \frac{2}{\pi(2p)!} \int_0^\infty dk D^{2p} \ln |f(k)| \quad (85)$$

$$d_p = \frac{2}{\pi(2p-1)!} \int_0^\infty dk D^{2p-1} \eta(k) \quad (86)$$

where a_p and d_p are the coefficients of the asymptotic expansion of $\eta(k)$ and $\Delta(k)$, Eqs. (33) and (34). Using Eq. (22), or rather its generalized version when bound states are present, which reads

$$|f(k)| = \left[\prod_n (1 - k^{-2} E_n) \right] \exp [\Delta(k)] \quad (87)$$

with $\Delta(k)$ always defined by Eq. (23), it is possible to recast Eq. (85) into the form

$$a_p = -\frac{2}{\pi(2p)!} \int_0^\infty dk D^{2p} \Delta(k) \quad (88)$$

Actually it is possible to derive these explicit expressions for the asymptotic coefficients a_n and d_n by more straightforward, if perhaps less rigorous, techniques [Roberts, 1964].

The results embodied in Eqs. (85) and (87) do not provide an explicit connection between the energy moments of the phase shifts [and the logarithm $\Delta(k)$ of the modulus of the Jost function], the bound-state energies, and the interaction. Such a connection is introduced using the explicit asymptotic expressions of the phase shift and the modulus of the Jost function, or of its logarithm, in terms of the potential. Such expressions are derived and discussed

in the following section, using a technique due to Verde [1955]. Similar results have been obtained with a different procedure by Buslaev and Faddeev [1960], which state that

$$a_p = (-)^p 2^{-2p-1} Q_{2p+1} \quad (89)$$

$$d_p = p^{-1} \sum_n (E_n)^p + (-)^{p+1} 2^{-2p} Q_{2p} \quad (90)$$

where the coefficients Q_n are obtained from the recursion relations

$$Q_p = \tilde{V}_{p-1} + \sum_{j=1}^{p-1} \left(\frac{j}{p}\right) \tilde{V}_{p-j-1} Q_j \quad (91)$$

where, by definition,

$$\tilde{V}_\ell = \lim_{r \rightarrow \infty} [\tilde{V}_\ell(r)] \quad (92)$$

and the functions $\tilde{V}_\ell(r)$ are obtained by recursion from

$$\tilde{V}_0(r) = - \int_0^r ds V(s) \quad (93)$$

and

$$\tilde{V}_\ell(r) = V^{(\ell-1)}(0) + \sum_{m=0}^{\ell-1} \binom{\ell-1}{m} \int_0^r ds \tilde{V}_m(s) V_{(s)}^{(\ell-1-m)} \quad (94)$$

As usual, $V^{(n)}(r)$ indicates the n th derivative of the potential, so that $V^{(n)}(0) = V_n$ are the quantities we are trying to compute. Note that, for $p = 1$, Eq. (91) yields

$$Q_1 = \tilde{V}_0 \quad (95)$$

The recursion relations (91) are solved by the formula [Calogero and Degasperis, 1968]

$$Q_p = \sum \left(-1 + \sum_{j=1}^{\ell} s_j \right)! \prod_{j=1}^{\ell} \frac{(V_{j-1})^{s_j}}{s_j!} \quad (96)$$

where the sum extends over all positive integral values of l and s_l and all nonnegative integral values of s_j ($j = 1, 2, \dots, l-1$) such that

$$\sum_{j=1}^l s_j = p \quad (97)$$

Explicit expressions of the first four coefficients Q_p are

$$Q_1 = - \int_0^\infty dr V(r) \quad (98)$$

$$Q_2 = V(0) \quad (99)$$

$$Q_3 = V'(0) + \int_0^\infty dr V^2(r) \quad (100)$$

$$Q_4 = V''(0) - 2V^2(0) \quad (101)$$

Inserting these expressions of Q_2 and Q_4 in Eqs. (90) and (86) *Buslaev and Faddeev* [1960] obtain explicit expressions for $V_0 = V(0)$ and $V_2 = V''(0)$ in terms of the generalized energy moments of the scattering phase shift given in Eq. (82). It is however clear that this approach cannot produce an explicit expression for the first derivative $V_1 = V'(0)$ of the potential at the origin, nor for the higher derivatives, except in the case of an even potential; for instance, it is easily seen that Q_6 is a linear combination of $V_4, V_2, V_0, (V_1)^2$, and $(V_0)^3$.

THE METHOD OF ASYMPTOTIC EXPANSIONS

This method is based on the relationship between the asymptotic expansion of the scattering parameters, and the interaction.

We begin with an analysis, following *Verde* [1955], of the asymptotic expansion at large k of the auxiliary function

$$g(k, r) = ik + \frac{f'(k, r)}{f(k, r)} \quad (102)$$

The radial Schrödinger equation satisfied by $f(k, r)$ implies for $g(k, r)$ the Riccati equation

$$g'(k, r) = V(r) + 2ikg(k, r) - g^2(k, r) \quad (103)$$

while the condition characterizing the Jost solution $f(k, r)$, Eq. (19), implies that $g(k, r)$ vanishes at large r :

$$\lim_{r \rightarrow \infty} [g(k, r)] = 0 \quad (104)$$

On the other hand, at the origin

$$g(k, 0) = g(k) \quad (105)$$

where the function $g(k)$ is the same one introduced in Eq. (24).

Introducing the asymptotic expansion at large k

$$g(k, r) = \sum_{n=0}^N g_n(r) (-2ik)^{-n-1} + O(k^{-N-2}) \quad (106)$$

into the Riccati Eq. (103), we obtain

$$g'_n(r) + g_{n+1}(r) + \sum_{m=0}^{n-1} g_m(r) g_{n-1-m}(r) = 0, \quad n \geq 1 \quad (107)$$

These recursive relations, together with the starting conditions

$$g_0(r) = V(r) \quad (108)$$

$$g_1(r) = -V'(r) \quad (109)$$

and the asymptotic property

$$g_n(\infty) = 0 \quad (110)$$

are sufficient to determine all the functions $g_n(r)$ in terms of the potential. Moreover, it is clear that $g_n(r)$ is an algebraic expression involving $V^{(p)}(r)$ for p up to n ; for instance,

$$g_2(r) = V''(r) - V^2(r) \quad (111)$$

$$g_3(r) = -V'''(r) - 4V(r)V'(r) \quad (112)$$

and so on. It is clear from the structure of these equations that they can be inverted, so as to obtain $V^{(p)}(r)$ as an algebraic expression involving the quantities $g_n(r)$ for n up to p ; and it is easily seen that the general structure of the resulting expression is

$$V^{(p)}(r) = \sum \left\{ p, \ell \mid s_1, s_2, \dots, s_\ell \right\} \prod_{j=1}^{\ell} g_{s_j}(r) \quad (113)$$

where the sum extends over all positive integers ℓ and over all sets of nonnegative integers s_j (a set being by definition independent from the order in which the indices s_j appear), and the coefficients $\{p, \ell \mid s_1, s_2, \dots, s_\ell\}$ vanish unless the equality

$$\sum_{j=1}^{\ell} s_j = p - 2\ell + 2 \quad (114)$$

is satisfied. This "dimensional" condition severely restricts the number of terms in the sum. For instance the allowed values of ℓ are restricted by

$$\ell \leq 1 + \frac{p}{2} \quad (115)$$

The coefficients $\{p, \ell \mid s_1, s_2, \dots, s_\ell\}$ are discussed in detail by *Calogero and Degasperis* [1963]; a table of all their values for p up to 10 is also given.

For $r = 0$ Eq. (113) becomes

$$V_p = \sum \left\{ p, \ell \mid s_1, s_2, \dots, s_\ell \right\} \prod_{j=1}^{\ell} g_{s_j} \quad (116)$$

where the quantities

$$g_s = g_s(0) \quad (117)$$

are the coefficients of the asymptotic expansion of $g(k)$:

$$g(k) = \sum_{n=0}^N g_n (-2ik)^{-n-1} + O(k^{-N-2}) \quad (118)$$

To realize the task stated in the title of this paper, it is now sufficient to express the quantities g_n in terms of the scattering phase shift. This can be easily done starting from the explicit expression of $g(k)$ in terms of the phase

shifts [Eqs. (25), (28), (22), and (23)] and exploiting the analyticity properties of $g(k)$, according to a technique first suggested, in a similar context, by *Roberts* [1964b]. We report here the results, relegating their proof to Appendix D. They consist of the formulas

$$g_{2p+1} = (-)^{p+1} \frac{2^{2p+3}}{\pi(2p+2)!} \int_0^\infty dk D^{2p+2} [1 - |f(k)|^{-2}] \quad (119)$$

$$= (-)^{p+1} \frac{2^{2p+3}}{(2p+3)!} \int_0^\infty dk D^{2p+3} [k^{-1} |f(k)|^{-2}] \quad (120)$$

$$g_{2p} = (-)^{p+1} 2^{2p+1} \sum \prod_{j=1}^{\ell} \frac{(-2d_j)^{s_j}}{s_j!} \quad (121)$$

the sum in the last equation being extended over all positive integral values of ℓ and s_ℓ and nonnegative integral values of s_j , ($j = 1, 2, \dots, \ell - 1$), such that

$$\sum_{j=1}^{\ell} j s_j = p + 1 \quad (122)$$

Note that this last condition restricts severely the range of allowed values of ℓ and of the indices s_j ; for instance, it implies

$$1 \leq \ell \leq 1 + p \quad (123)$$

The quantities d_j in Eq. (121) are proportional to the moments of the scattering phase shift, being defined by

$$d_p = \frac{2}{\pi(2p-1)!} \int_0^\infty dk D^{2p-1} \eta(k) \quad (124)$$

and they coincide with the coefficients of the asymptotic expansion of $\Delta(k)$ [Eq. (34)]. The operator D entering in this equation and in Eqs. (119–120) is the same one discussed in Appendix A, namely,

$$D = k \frac{d}{dk} k \quad (125)$$

Note that Eq. (120) obtains from (119) performing one integration by parts (see also Appendix A).

We thus see that the parameters g_n of the asymptotic expansion (118) of $g(k)$ can be expressed explicitly in terms of the scattering phase shift; the parameters g_n with n even are algebraic combinations of the energy moments (124) of the scattering phase shift, while the parameters g_n with n odd are directly proportional to the generalized moments of the spectral function $1 - |f(k)|^{-2}$ [and this function can be obtained from the scattering phase shift through the explicit formulas (22) and (23)]. Inserting these expressions in Eq. (116), we obtain the final expression

$$V_p = \sum_{\ell} [p, \ell | s_2, s_3, \dots, s_{\ell+1}] \prod_{j=2}^{\ell+1} (J_j)^{s_j} \quad (126)$$

where

$$J_{2j} = (-)^{j+1} \frac{2^{2j+3}}{\pi(2j-1)!} \int_0^\infty dk D^{2j-1} \eta(k) \quad (127)$$

$$= (-)^j 2^{2j+2} d_j \quad (128)$$

and

$$J_{2j+1} = (-)^j \frac{2^{2j+1}}{\pi(2j+1)!} \int_0^\infty dk D^{2j+1} [k^{-1} |f(k)|^{-2}] \quad (129)$$

$$= g_{2j-1} \quad (130)$$

The sum in Eq. (126) extends over all positive integral values of ℓ and $s_{\ell+1}$ and all nonnegative integral values of s_j ($j = 2, 3, \dots, \ell$), such that

$$\sum_{j=2}^{\ell+1} j s_j = p + 2 \quad (131)$$

Thus the maximum value of ℓ appearing in (126) is $p + 1$. The coefficients $[p, \ell | s_2, s_3, \dots, s_{\ell+1}]$ are related in an obvious way, through Eqs. (119-121) and (127-130), to the coefficients $\{p, \ell | s_1, s_2, \dots, s_\ell\}$ of Eq. (116). They have been computed by Calogero and Degasperis [1968] for all values of p up to 10, and are given in Table I of their paper. Using these coefficients, it is possible to write the values at the origin of the potential and its first 10 derivatives, in terms of the phase shift. These explicit expressions are given in the following section, which contains a general summary of the results for the case of S -wave scattering on a holomorphic potential, including the possibility that bound states are present.

Finally, we note that while the general philosophy of the approach given in this section was introduced by Roberts [1964b], his method could not yield all the results reported here, because it focused on the asymptotic expansion of the function

$$h(k) = - \int_0^\infty dr \frac{d}{dk} g(k, r) \quad (132)$$

rather than on the function $g(k, r)$. As explained in detail by *Calogero and Degasperis* [1968] (see, in particular, their Appendix H), the information derivable from a study of the asymptotic expansion of $h(k)$ corresponds as far as the values of the potential and its derivatives at the origin are concerned, only to that derivable from an expansion of the odd part of $g(k)$, namely that transmitted by the coefficients g_n with n even [see Eq. (118)]. Thus *Roberts* [1946b] could not obtain an expression for $V'(0) = V_1$, nor could he obtain expressions for V_p for $p > 2$, except in the special case of an even potential—that is, a potential such that all V_p with odd p vanish.

S-WAVE SCATTERING ON A HOLOMORPHIC POTENTIAL: SUMMARY OF RESULTS

Let $\eta(k)$ be the S -wave scattering phase shift produced by the potential $V(r)$, E_n the (negative) energies of the corresponding (S -wave) bound states, and C_n the normalization coefficients of these bound states [defined by Eqs. (15) and (16), with $\ell = 0$].

Let V_p be the value at the origin of the p th derivative of the potential (assumed to be holomorphic at $r = 0$ and to satisfy the usual requirements in scattering theory, detailed earlier):

$$V(r) = \sum_{p=0}^{\infty} \frac{V_p r^p}{p!} \quad (133)$$

Then the following formula holds:

$$V_p = \sum [p, \ell | s_2, s_3, \dots, s_{\ell+1}] \prod_{j=2}^{\ell+1} (J_j)^{s_j} \quad (134)$$

the sum being extended over all positive integral values of ℓ and $s_{\ell+1}$ and nonnegative integral values of s_j that are consistent with the dimensional condition

$$\sum_{j=2}^{\ell+1} j s_j = p + 2 \quad (135)$$

The universal coefficients $[p, \ell | s_2, s_3, \dots, s_{\ell+1}]$ depend only on the indices shown; procedures to compute them, together with a table of all the nonvanishing coefficients for p up to 10 [see Eqs. (145) through (155)] are given by *Calogero and Degasperis* [1968]. The integrals J_n with even n are essentially the generalized momentum moment (of odd order $n - 1$) of the scattering phase shift; they depend on the phase shift $\eta(k)$ and on the bound state energies E_n , but not on the normalization coefficients C_n :

$$J_{2j} = (-)^{j+1} 2^{2j+2} \left\{ \frac{1}{j} \sum_n (E_n)^j + \frac{2}{\pi(2j-1)!} \int_0^\infty dk k^{2j-1} \frac{d^{2j-1}}{dk^{2j-1}} [k^{2j-1} \eta(k)] \right\} \quad (136)$$

$$= (-)^{j+1} 2^{2j+2} \left\{ \frac{1}{j} \sum_n (E_n)^j - \frac{2}{\pi} \int_0^\infty dk k^{2j-1} \left[\eta(k) - \sum_{n=0}^{j-1} a_n k^{-2n-1} \right] \right\} \quad (137)$$

The coefficients a_n entering in Eq. (137) are those characterizing the asymptotic behavior of the phase shift

$$\eta(k) = \sum_{n=0}^N a_n k^{-2n-1} + O(k^{-2N-3}) \quad (138)$$

They can be also defined by the integral expressions

$$a_n = \frac{1}{(2n)!} \int_0^\infty dk k^{2n} \frac{d^{2n+1}}{dk^{2n+1}} [k^{2n+1} \eta(k)] \quad (139)$$

The integrals J_n with odd n are essentially the generalized momentum moments (or order $n-1$) of the spectral function $1 - |f(k)|^{-2}$; they depend not only on the phase shift and the bound-state energies, but also on the bound-state normalization coefficients C_n , defined by Eq. (15):

$$J_{2j+1} = (-)^{j+1} 2^{2j} \left\{ \sum_n C_n (E_n)^{j-1} - \frac{2}{\pi(2j+1)!} \int_0^\infty dk k^{2j+1} \frac{d^{2j+1}}{dk^{2j+1}} [k^{2j+1} |f(k)|^{-2}] \right\} \quad (140)$$

$$= (-)^{j+1} 2^{2j} \left\{ \sum_n C_n (E_n)^{j-1} - \frac{2}{\pi} \int_0^\infty dk k^{2j} \left[1 - |f(k)|^{-2} + \sum_{n=0}^{j-1} (-)^n \tilde{g}_{2n} k^{-2n-2} \right] \right\} \quad (141)$$

The coefficients $(-)^n \tilde{g}_{2n}$ entering in Eq. (141) are those characterizing the asymptotic behavior of the spectral function

$$1 - |f(k)|^{-2} = - \sum_{n=0}^N (-)^n \tilde{g}_{2n} k^{-2n-2} + O(k^{-2N-4}) \quad (142)$$

The spectral function is expressed in terms of the phase shift and the bound-state energies through the formula

$$|f(k)| = \exp [\Delta(k)] \prod_n \left(1 - \frac{E_n}{k^2}\right) \quad (143)$$

where

$$\Delta(k) = -\frac{2}{\pi} P \int_0^\infty dq \frac{q}{(q^2 - k^2)} \eta(q) \quad (144)$$

Below we display the explicit expressions for the values at the origin of the first 10 derivatives of the potential, implied by Eq. (134) and the values of the coefficients $\{p, \ell | s_2, s_3, \dots, s_{\ell+1}\}$ given by *Calogero and Degasperis* [1968]:

$$V_0 = -\frac{1}{4} J_2 \quad (145)$$

$$V_1 = -J_3 \quad (146)$$

$$V_2 = \frac{1}{8} (J_2)^2 - \frac{1}{4} J_4 \quad (147)$$

$$V_3 = J_2 J_3 - J_5 \quad (148)$$

$$V_4 = -\frac{1}{6} (J_2)^3 + 5(J_3)^2 + \frac{1}{2} J_2 J_4 - \frac{1}{4} J_6 \quad (149)$$

$$V_5 = -\frac{13}{4} (J_2)^2 J_3 + \frac{9}{2} J_3 J_4 + 2J_2 J_5 - J_7 \quad (150)$$

$$V_6 = \frac{1}{2} (J_2)^4 - 28J_2(J_3)^2 + \frac{5}{4} (J_4)^2 - 2(J_2)^2 J_4 + 28J_3 J_5 + \frac{3}{4} J_2 J_6 - \frac{1}{4} J_8 \quad (151)$$

$$V_7 = -140(J_3)^3 + \frac{97}{6} (J_2)^3 J_3 - 37J_2 J_3 J_4 + 17J_4 J_5 - \frac{23}{2} (J_2)^2 J_5 + 10J_3 J_6 + 3J_2 J_7 - J_9 \quad (152)$$

$$V_8 = -\frac{38}{15} (J_2)^5 + 258(J_2)^2 (J_3)^2 - 14J_2 (J_4)^2 + \frac{38}{3} (J_2)^3 J_4 + 69(J_5)^2 - 246J_2 J_3 J_5 \\ + 7J_4 J_6 - 5(J_2)^2 J_6 + 54J_3 J_7 + J_2 J_8 - \frac{1}{4} J_{10} \quad (153)$$

$$V_9 = 2310J_2(J_3)^3 - \frac{6047}{48}(J_2)^4J_3 - \frac{791}{4}J_3(J_4)^2 + \frac{1667}{4}(J_2)^2J_3J_4 - 208J_2J_4J_5 - 2310(J_3)^2J_5 + \frac{523}{6}(J_2)^3J_5 \\ + \frac{125}{2}J_3J_6 - 115J_2J_3J_6 + \frac{83}{2}J_4J_7 - \frac{107}{4}(J_2)^2J_7 + \frac{35}{2}J_3J_8 + 4J_2J_9 - J_{11} \quad (154)$$

$$V_{10} = \frac{59}{3}(J_2)^6 + 11550(J_3)^4 + \frac{9196}{3}(J_2)^3(J_3)^2 - \frac{164}{3}(J_4)^3 + 190(J_2)^2(J_4)^2 + 5852J_2(J_3)^2J_4 \\ - 836J_2(J_5)^2 - 3652J_3J_4J_5 + 3344(J_2)^2J_3J_5 + \frac{63}{4}(J_6)^2 - 108J_2J_4J_6 - 1100(J_3)^2J_6 + 46(J_2)^3J_6 \\ + 418J_5J_7 - 628J_2J_3J_7 + 15J_4J_8 - 10(J_2)^2J_8 + 88J_3J_9 + \frac{5}{4}J_2J_{10} - \frac{1}{4}J_{12} \quad (155)$$

In the very special case of an even potential — a potential such that $V_n = 0$ for n odd — all the integrals J_n with n odd must vanish; this implies an (infinite) number of constraints that must be satisfied by the phase shift, the bound-state energies, and the normalization coefficients [Eqs. (140) and (141)]. Thus the binding energies and normalization coefficients of an even potential are determined by the scattering phase shift, which also must satisfy an infinite number of integral conditions. For instance, if only one bound state is present, and the potential is even, the normalization coefficient C of the bound state is given by the explicit equation

$$C = \frac{1}{3\pi} \int_0^\infty dk k^3 \frac{d^2}{dk^2} \left\{ k^6 (k^2 - E)^{-2} \exp[-2\Delta(k)] \right\} \quad (156)$$

while the energy E of the bound state can be obtained from the equation

$$E = \frac{1}{20} \left\{ \int_0^\infty dk k^5 \frac{d^5}{dk^5} k^8 (k^2 - E)^{-2} \exp[-2\Delta(k)] \right\} \\ \left\{ \int_0^\infty dk k^3 \frac{d^3}{dk^3} k^6 (k^2 - E)^{-2} \exp[-2\Delta(k)] \right\}^{-1} \quad (157)$$

Here $\Delta(k)$ is given in terms of $\eta(k)$ by Eq. (144). Of course, for an even potential the (even) derivatives of the potential at the origin are expressed as polynomials of the integrals J_n with n even [Eqs. (136) and (137)] — that is, only in terms of the generalized momentum moments of the phase shift.

The possibility of expressing all the derivatives of the potential at the origin in the form (134) has some implications for the so called "Jost equivalent potentials" (which have the same phase shift and bound-state energies and therefore the same Jost function $f(k) = |f(k)| \exp[i\eta(k)]$ [see Eqs. (143), (144)], but possibly yielding different normalization coefficients C_n) and for the "Bargmann potentials" (whose Jost function is a polynomial in k). For these, and a brief discussion of the possibility of finding other explicit relationships between the scattering data and the potential, see Calogero and Degasperis [1968]; see also the last section of this paper.

RESULTS FOR HIGHER PARTIAL WAVES

Some of the approaches described in the previous sections extend trivially to the case of higher partial waves. Indeed the result that is at the origin of this line of research, namely the explicit expression for the value V_0 of the potential at the origin in terms of the scattering phase shift, was originally given by *Newton* [1956] and, less explicitly, by *Faddeev* [1957] for an arbitrary partial wave:

$$V_0 = - \frac{4}{2\ell + 1} \sum_n E_{\ell n} - \frac{8}{\pi(2\ell + 1)} \int_0^\infty dk k \frac{d}{dk} [k\eta_\ell(k)] \quad (158)$$

where $\eta_\ell(k)$ is the ℓ -wave scattering phase shift and $E_{\ell n}$ are the energies of the bound states (if any) with angular momentum ℓ . The convergence of the integral in the right-hand side is guaranteed by the asymptotic behavior of the phase shift

$$\eta_\ell(k) \xrightarrow[k \rightarrow \infty]{} a_{\ell 0} k^{-1} + O(k^{-3} \ln k) \quad (159)$$

which implies

$$\frac{d}{dk} [k\eta_\ell(k)] \xrightarrow[k \rightarrow \infty]{} O(k^{-3} \ln k) \quad (160)$$

As we discuss below, this asymptotic behavior is related to the finiteness of the value of the potential at the origin. The trace method discussed earlier has been extended to higher partial waves by *Percival and Roberts* [1963], following the route introduced for S -waves by *Percival* [1962], and by *Buslaev* [1962, 1966], who based his approach on the method introduced for S waves by *Buslaev and Faddeev* [1960] but was actually able to treat the scattering problem directly in three-dimensional space, without performing a partial wave expansion. However, as noted, the trace method is unable to yield the values of all the derivatives of the potential at the origin in terms of the scattering and bound-state data, except in the very special case of an even potential — that is, only if it is a priori known that all the derivatives of odd order of the potential vanish at the origin.

The main difficulty in extending the method of asymptotic expansions, which in the S -wave case provided the most complete solution of our problem, to the case of higher partial waves and also to the case of nonholomorphic potentials is due to the fact that the asymptotic expansion of the scattering phase shift at large k contains generally logarithmic terms in addition to powers if the restriction to S waves (and to holomorphic potentials) is abandoned. There is, however, a subclass of holomorphic potentials such that the corresponding scattering phase shifts $\eta_\ell(k)$, for all $\ell \leq L$, still contain only (inverse) powers of k in their asymptotic expansions. The condition characterizing this class of potentials is the requirement that the values at the origin of the first L derivatives of odd order of the potential vanish:

$$V_{2n+1} = 0 \quad \text{for } n = 0, 1, \dots, L-1 \quad (161)$$

Note that there is clearly no restriction for $L = 0$, and that these conditions are automatically satisfied for all values of L if the potential is even.

The discovery of this property is due to *Corbella* [1971], who showed that, for this special class of potentials, it is possible to reduce the ℓ -wave problem, with $\ell \leq L$, to an S -wave problem with a new potential that is still holomorphic at the origin. The process of reduction implies that the S -wave phase shift and bound-state parameters of the new problem essentially coincide with the ℓ -wave phase shift and bound-state parameters of the original problem (see below). Then, of course, the results of *Calogero and Degasperis* [1968] for the S -wave case can be taken over, and, in principle, the complete solution of our problem is achieved for this special class of potentials. We refer for a detailed treatment to the paper by *Corbella* [1971], limiting discussion here to a simple outline of the method and results. The original mathematical trick for the transition from a P -wave problem ($\ell = 1$) to an S -wave one ($\ell = 0$) is due to *Calogero* [unpublished, 1970]; its development and use is due to *Corbella* [1971].

The basis of the method is as follows. Let $y(r)$ be the solution of the differential equation

$$y''(r) + [k^2 - W(r)]y(r) = 0 \quad (162)$$

and introduce a function $f(r)$ through the Riccati equation

$$f'(r) = f^2(r) - W(r) \quad (163)$$

Define the function $\tilde{y}(r)$ through

$$\tilde{y}(r) = f(r)y(r) + y'(r) \quad (164)$$

Then this function satisfies the new differential equation

$$\tilde{y}''(r) + [k^2 - \tilde{W}(r)]\tilde{y}(r) = 0 \quad (165)$$

with

$$\tilde{W}(r) = W(r) + 2f'(r) \quad (166)$$

as can be verified by direct substitution. Assume now that

$$W(r) = \ell(\ell + 1)r^{-2} + V(r) \quad (167)$$

so that Eq. (162) becomes the ℓ -wave radial Schrödinger equation

$$\phi_\ell'(r) + [k^2 - \ell(\ell + 1)r^{-2} - V(r)]\phi_\ell(r) = 0 \quad (168)$$

Then set

$$f_\ell(r) = \frac{\ell}{r} + g_\ell(r) \quad (169)$$

so that Eq. (163) becomes

$$g'_\ell(r) = g_\ell^2(r) + 2 \frac{\ell}{r} g_\ell(r) - V(r) \quad (170)$$

while Eq. (164) defines a new function

$$\tilde{\phi}_{\ell-1}(r) = \left[\frac{\ell}{r} + g_\ell(r) \right] \phi_\ell(r) + \phi'_\ell(r) \quad (171)$$

The index $\ell - 1$ has been introduced because Eq. (165) now reads

$$\tilde{\phi}_{\ell-1}''(r) + \left[k^2 - (\ell-1)\ell r^{-2} - \tilde{V}(r) \right] \tilde{\phi}_{\ell-1}(r) = 0 \quad (172)$$

with

$$\tilde{V}(r) = V(r) - 2 g'_\ell(r) \quad (173)$$

Thus the function $\tilde{\phi}_{\ell-1}(r)$ satisfies now the radial Schrödinger equation appropriate to the angular momentum quantum number $\ell - 1$, with the new potential $\tilde{V}(r)$ of Eq. (173), the function $g_\ell(r)$ being defined by the Riccati equation (170).

Repeating this procedure ℓ times, one arrives at the equation

$$\tilde{\tilde{\phi}}(r) + \left[k^2 - \tilde{\tilde{V}}(r) \right] \tilde{\tilde{\phi}}(r) = 0 \quad (174)$$

with

$$\tilde{\tilde{V}}(r) = V(r) - 2 \sum_{n=1}^{\ell} g'_n(r) \quad (175)$$

The functions $g_n(r)$ are defined by the recursive Riccati equations

$$g'_\ell(r) = g_\ell^2(r) + 2 \frac{\ell}{r} g_\ell(r) - V(r) \quad (176)$$

$$g'_n(r) = g_n^2(r) + 2 \frac{n}{r} g_n(r) - V(r) - 2 \sum_{m=n+1}^{\ell} g'_m(r), \quad n = \ell-1, \ell-2, \dots, 1 \quad (177)$$

Equation (174) is the S -wave radial Schrödinger equation with the potential $\tilde{\tilde{V}}(r)$, and it is easy to see that the S -wave phase shift and bound-state parameters corresponding to this potential coincide, or are very simply related.

to the ℓ -wave phase shifts and bound-state parameters produced by the original potential $V(r)$. Indeed, the explicit relations given by Corbella [1971] are

$$\eta_{\ell}(k) = \tilde{\eta}_{\ell}(k) \quad (178)$$

$$E_{\ell n} = \tilde{E}_n \quad (179)$$

$$(-E_{\ell n})^{-\ell} C_{\ell n} = \tilde{C}_n \quad (180)$$

where the quantities $\eta_{\ell}(k)$, $E_{\ell n}$, and $C_{\ell n}$ are the ℓ -wave phase shifts, bound-state energies, and bound-state normalization coefficients for the original problem with potential $V(r)$, while $\tilde{\eta}_{\ell}(k)$, \tilde{E}_n and \tilde{C}_n are the S -wave phase shift, bound-state energies and normalization coefficients for the potential $\tilde{V}(r)$.

It is easy to prove [Corbella, 1971] that if the condition (161) holds and if the original potential $V(r)$ is holomorphic at $r=0$, the final potential $\tilde{V}(r)$ of Eq. (175) also is holomorphic at $r=0$ (for $\ell \leq L$). Then the results of Calogero and Degasperis [1968] are applicable to the S -wave problem with the potential $\tilde{V}(r)$, so that the values at the origin of this potential and its derivatives are expressible by the explicit formulas (145) - (155) in terms of the phase shift $\tilde{\eta}_{\ell}(k)$ and the bound-state data \tilde{E}_n and \tilde{C}_n , which are very simply related to the corresponding quantities for the original ℓ -wave problem with potential $V(r)$ [see Eqs. (178) through (180)]. But the definition (175), together with the recursive equations (176), (177), imply definite relationships between the values at the origin of the potential $\tilde{V}(r)$ and its derivatives, and the values at the origin of the potential $V(r)$ and its derivatives. This is the case only for values of ℓ such that $\ell \leq L$, with L defined by the condition (161). In particular Corbella [1971] gives the explicit formulas

$$V_0 = \frac{\tilde{V}_0}{2\ell + 1} \quad (181)$$

$$V_2 = -3 \left\{ \frac{\tilde{V}_2 - [2\ell/(2\ell + 1)](\tilde{V}_0)^2}{(2\ell - 1)(2\ell + 1)(2\ell + 3)} \right\} \quad (182)$$

$$V_3 = -\frac{1}{3} \left\{ \tilde{V}_3 - \frac{2}{3} \tilde{V}_0 \tilde{V}_1 \right\} \quad \ell = 1 \quad (183)$$

$$V_4 = 45 \left\{ \frac{\tilde{V}_4 - [4\ell/(6\ell + 3)] \tilde{V}_0 \tilde{V}_2 + [4\ell(2\ell - 1)/(6\ell + 3)^2] (\tilde{V}_0)^3 - (5/24)(\tilde{V}_1)^2}{(2\ell - 3)(2\ell - 1)(2\ell + 1)(2\ell + 3)(2\ell + 5)} \right\} \quad (184)$$

$$V_5 = -\frac{1}{2} \left\{ \tilde{V}_5 - \frac{1}{3} \tilde{V}_0 \tilde{V}_3 - \frac{3}{10} \tilde{V}_1 \tilde{V}_2 + \frac{4}{45} (\tilde{V}_0)^2 \tilde{V}_1 \right\} \quad \ell = 1 \quad (185)$$

$$V_5 = \frac{1}{10} \left\{ \tilde{V}_5 - \frac{2}{5} \tilde{V}_0 \tilde{V}_3 - \frac{3}{10} \tilde{V}_1 \tilde{V}_2 + \frac{2}{15} (\tilde{V}_0)^2 \tilde{V}_1 \right\} \quad \ell = 2 \quad (186)$$

where, of course, V_n is the n th derivative of the potential $V(r)$ evaluated at the origin, and \tilde{V}_n is the n th derivative of the potential $\tilde{V}(r)$, also evaluated at the origin. The validity of the restriction (161) is always assumed. These equations, together with the correspondence of Eqs. (178 – 180) and the results of *Calogero and Degasperis* [1968] [see Eqs. (145 – 155)], yield explicit expressions of V_n , for n up to 5, in terms of the ℓ -wave scattering and bound-state data. Note that the first of these equations reproduces the original result of *Faddeev* [1957] and *Newton* [1956], Eq. (158).

It must be emphasized that the explicit expressions of the potential and its derivatives at the origin in terms of phase shifts and bound-state data for different angular momenta imply definite integral relationships between scattering phase shifts and bound-state data corresponding to different angular momenta. One consequence of these relationships is the possibility of expressing the bound-state parameters in terms of the scattering phase shifts. For instance for a potential that possesses only one S -wave bound state and no P -wave bound state the Newton-Faddeev Eq. (158) immediately implies

$$E = \frac{2}{\pi} \int_0^\infty dk \frac{d}{dk} \left\{ k \left[\eta_0(k) - \frac{1}{3} \eta_1(k) \right] \right\} \quad (187)$$

where E is the energy of the S -wave bound state, and $\eta_0(k)$ resp. $\eta_1(k)$ are the S -wave resp. P -wave scattering phase shifts.

RESULTS FOR THE KLEIN-GORDON AND DIRAC EQUATIONS

The results of *Calogero and Degasperis* [1968] have been extended to the case of the Klein-Gordon equation by *Degasperis* [1970], and to the Dirac equation by *Corbella* [1970]. The interested reader is referred to these two papers for a clear and detailed presentation, which cannot be given here as it requires a previous treatment of the Klein-Gordon and Dirac inverse scattering problems. The results obtained are interesting per se and constitute progress toward a more complete solution of the inverse scattering problem for Klein-Gordon and Dirac equations than had been previously achieved.

The scattering and bound-state problems for particles described by the Klein-Gordon or the Dirac equations present some new features, connected with the relativistic nature of these equations. The most important is the appearance of negative energy solutions, which are interpreted as corresponding to antiparticles. Another characteristic feature is the possibility of generalizing the basic equations introducing two potentials, characterized by different transformation properties — one behaving as the fourth component of a relativistic fourvector and the other as a relativistic scalar. The canonical forms of the Klein-Gordon and Dirac equations, however, contain only the former.

The inverse scattering problem that has been solved requires as input data the scattering phase shift, bound-state energies, and normalization parameters for one value of the angular momentum, *both for particles and antiparticles*, it yields as output (through the solution of integral equations) two potential functions. The requirement that one of these vanish identically, which must be imposed if the usual form of the Klein-Gordon or Dirac equation is being considered, clearly implies restrictions on the input; the particle and antiparticle data are then not independent, and indeed it is plausible to conjecture that the data for the particle (or the antiparticle) are sufficient to determine the potential, and therefore also to determine the data for the antiparticle (or for the particle). However, no definite results have been proved so far.

The results obtained by *Degasperis* [1970] and *Corbella* [1970] are explicit expressions for the values at the origin of the two potentials, and their first few derivatives, in terms of the phase shifts and bound-state data for the particle and the antiparticle (for S waves in the Klein-Gordon case, for S and P waves in the Dirac case). Then the requirement that only one potential be present — that the second one vanish identically — provides explicit integral relationships between the scattering and bound-state data for the particle and for the antiparticle. No such relationships had been previously obtained, and they constitute a first step towards a more complete understanding of the inverse scattering problem for the usual Klein-Gordon and Dirac equations. Indeed, if all the derivatives of the potentials at the origin are considered, the requirement that all vanish for the second potential yields an infinite set of integral relations between the scattering and bound-state data for the particle and those for the antiparticle, which might be sufficient to determine one set from the other. This suggests a procedure to prove the conjecture mentioned above.

FINAL REMARKS

We have reviewed the results obtained for a rather specific problem — that of obtaining, in the framework of quantal potential scattering, *explicit* expressions for the values at the origin of the potential and its derivatives, from a knowledge of the scattering and bound-state data (for one value of the angular momentum). In this final section we briefly discuss the phenomenological applicability of these results. We indicate the extensions that are naturally suggested, but have not yet been accomplished, and we mention some other developments that fit into the same philosophy of obtaining more explicit results than those yielded by the standard solution of the inverse scattering problem (that requires the solution of an integral equation).

The main obstacle to the phenomenological applicability of the results reported in this paper is the need to know the values of one scattering phase shift for all values of the energy. This difficulty is common, of course, to all approaches to the inverse scattering problem that take as input the data for a given angular momentum. It is, however, particularly relevant to the approach discussed in this paper, which focusses on the values at the origin of the potential and its derivatives, because these features of the potential are more sensitive to the information contained in the behavior of the potential at high energy. Indeed, in many phenomenological contexts, the high-energy behavior of the phase shift and the short range behavior of the potential are just the regimes where the representation of the physical problem in terms of a potential is less reliable. This same argument can also be used to justify the relevance of this approach; for the explicit relationships that we have discussed constitute the more appropriate way to evaluate to what extent the phenomenological data determine the potential, and its derivatives, at the origin, and thus they provide a quantitative basis to assess the consistency with phenomenology, and the definiteness, of the description of the interaction in terms of a potential, in the neighborhood of the origin. A useful tool in this connection is the possibility of obtaining the parameters characterizing the asymptotic behavior of the phase shifts from their overall dependence on the linear momentum k [see Eqs. (138 – 139)].

The more complete results have been obtained so far for the case of nonrelativistic S -wave scattering on a holomorphic potential, as discussed earlier. The extension of these results to potentials that have an r^{-1} singularity at the origin (more specifically, potentials such that $rV(r)$ is holomorphic at the origin) is closely connected with the generalization to higher partial waves [for potentials that do not obey the restriction of Eq. (161)]. A first result in this direction [*Degasperis*, private communication, 1967] notes that the Faddeev-Newton relation (158) remains essentially valid for a potential whose expansion at the origin has the form

$$V(r) = \frac{V_{-1}}{r} + \sum_{n=0}^{\infty} V_n \frac{r^n}{n!} \quad (188)$$

for it can then be written in the form

$$V_O = \bar{V}_O + \frac{8}{\pi} \int_0^{\infty} dk k [\eta(k) - \bar{\eta}(k)] \quad (189)$$

provided $\eta(k) - \bar{\eta}(k)$ vanishes at infinity sufficiently fast for the integral to converge. Since for potentials of type (188) the two leading terms in the asymptotic behavior of the phase shift are given by the asymptotic formula

$$\eta(k) = \frac{b_O \ln k}{k} + \frac{a_O}{k} + O\left(\frac{\ln k}{k^3}\right) \quad (190)$$

we can rewrite Eq. (189) in the form

$$V_O - \frac{8}{\pi} \int_0^{\infty} dk [k\eta(k) - b_O \ln k - a_O] = B_O(a_O, b_O) \quad (191)$$

where B_O is now an universal function of the arguments shown. An equivalent way to write this formula is

$$V_O - \frac{8}{\pi} \int_0^{\infty} dk \left[-k \frac{d}{dk} \right]^p [k\eta(k)] = B_O(a_O, b_O) \quad p = 2, 3, 4, \dots \quad (192)$$

The choice of an arbitrary integral value of p in Eq. (192) is justified by the remark that, if a function vanishes sufficiently fast at $k=0$ and as $k \rightarrow \infty$, then by partial integration

$$\int_0^{\infty} dk \left[-k \frac{d}{dk} \right] f(k) = \int_0^{\infty} dk f(k)$$

On the other hand, the need for the integer p in Eq. (192) to be larger than unity is connected with the asymptotic behavior (190) of the phase shift.

The universal function $B_O(a_O, b_O)$ has been recently determined by A. Degasperis [private communication, 1971]. Thus, the final formula for V_O is

$$V_O = \frac{8}{\pi} \int_0^{\infty} dk \left[-k \frac{d}{dk} \right]^p [k\eta(k)] - \frac{1}{3} \pi^2 b_O^2 \quad p = 2, 3, 4, \dots, \quad (193)$$

where b_0 is defined by Eq. (192) and is related to V_{-1} by the simple equation

$$V_{-1} = -2 b_0 \quad (194)$$

These equations have been written for the S -wave case without bound states. The inclusion of bound states presents no problem; nor does the extension to higher partial waves. Indeed, these recent results of Degasperis, together with the reduction technique discussed earlier, open up the possibility to obtain explicit expressions for all the derivatives of the potential, evaluated at the origin, in terms of the scattering phase shifts and bound-state data for higher partial waves, even if the potential does not satisfy the condition (161).

The possibility of treating potentials of type (188) and higher partial waves, besides being interesting per se, has a large potential for application; for instance, it opens the possibility of dealing with the nucleon-nucleon scattering problem, especially in the context of the so-called one-boson-exchange potentials [Calogero, 1969]. (Note that the formula for V_0 reported in this reference is incorrect, as it was obtained using the conjecture that the universal function B_0 vanishes when no bound states are present.)

Another generalization that would considerably widen the phenomenological applicability of the approach is to multichannel potential scattering problems. The only result for this case is contained in the original paper by Faddeev [1957]; it relates the value at the origin of the trace of the potential matrix to the generalized first momentum moment of the trace of the logarithm of the scattering matrix (or equivalently, of the sum of all the scattering eigen phase shifts).

As mentioned above, the explicit expressions for the values at the origin of the potential and its derivatives discussed in this paper are not the only explicit relations between the potential and the scattering and bound-state data that are known. Other explicit expressions have been obtained from a study of the asymptotic behavior of the scattering phase shift, which yields, for example, the formulas [Calogero and Degasperis, 1968]

$$\int_0^\infty dr V(r) = -2a_0 \quad (195)$$

$$\int_0^\infty dr V^2(r) = -8a_1 + 4 \sum_n C_n + \left[\frac{4}{(3\pi)} \right] \times \int_0^\infty dk k^3 \frac{d}{dk^3} \left[k^2 |f(k)|^{-2} \right] \quad (196)$$

where the coefficients a_n are those characterizing the asymptotic behavior of the S -wave scattering phase shift $\eta(k)$ through Eq. (138) (and are therefore expressed explicitly in terms of $\eta(k)$ by Eq. (139)), the modulus of the Jost function $|f(k)|$ is explicitly expressed in terms of the phase shift and the bound-state data by Eqs. (143-144), and C_n are the normalization constants of the S -wave bound states (defined by Eq. (15)). However, as explained by Calogero and Degasperis [1968], it is not possible to obtain other simple expressions of this kind; in fact, the next expression of this type that could be explicitly related to the scattering and bound-state data has the form

$$\int_0^\infty dr V^3(r) + \frac{1}{2} \int_0^\infty dr [V'(r)]^2 \quad (197)$$

namely, it contains the potential in a rather complicated way.

Other relations between the potential and the scattering and bound-state data, in which energy moments of the phase shifts of negative order enter, have been obtained by *Roberts* [1964a]. These relationships, however, do not involve the potential in an explicit way.

Another relevant result given by *Calogero et al* [1968] equates a simple functional of the potential to an explicit variational expression in terms of the scattering and bound state data. A simple consequence of this result is the following variational bound for the Laplace transform of the potential

$$\int_0^{\infty} dr V(r) \exp(-2pr) > -\sum_n C_n/(p^2 + E_n) + \frac{2}{\pi} \int_0^{\infty} dq [q^2/(q^2 + p^2)] [1 - |f(q)|^{-2}] \quad (198)$$

where the only restriction on the positive constant p is

$$p > (-E_n)^{1/2} \quad (199)$$

and the other symbols have the usual meaning (see above).

An appealing direction for research, consistent with the philosophy of searching for *explicit* relationships between the potential and the scattering and bound state data, but implying more radical modifications in the approach, takes as point of departure a kind of information different from that contained within a single scattering phase shift (assumed known for all energies) and the corresponding bound-state parameters. An instance of this kind is provided by the results of *Buslaev* [1962, 1966], who worked in terms of the full scattering amplitude (assumed known for all energies), although, of course, as mentioned in the first section, for a spherically symmetrical potential once the full scattering amplitude is known, the partial wave scattering amplitudes $\exp[i\eta_l(k)] \sin[\eta_l(k)]$ and therefore the scattering phase shifts $\eta_l(k)$ also are known.

Finally, a very interesting direction of research is the attempt to apply the techniques described here to the quantal three-body and many-body problems. Practically no results for the many-body problem have been obtained so far, besides those of *Uhlenbeck and Beth* [1936], consisting of an exploitation of the trace approach in the context of statistical mechanics (see also the more recent and complete contribution by *Berezin* [1964]). Results for the three-body problem that appear very interesting have been recently announced by *Buslaev and Merkurev* [1970].

APPENDIX A

In this appendix (that coincides with Appendix D of the paper by *Calogero and Degasperis* [1968]), we collect some useful operator and integral formulas.

First of all we introduce the operator

$$D = k \frac{d}{dk} k \quad (A1)$$

and we note the identity

$$D^n \equiv \left(k \frac{d}{dk} k \right)^n = k^n \left(\frac{d^n}{dk^n} \right) k^n \quad (A2)$$

which is easily proved by induction.

A remarkable property of the operator D is expressed by the following formula:

$$D^n \sum_{m=1}^N f_m k^{-m} = 0, \quad \text{if } n \geq N, \quad (A3a)$$

$$= (-)^n \sum_{m=n+1}^N f_m \frac{(m-1)!}{(m-n-1)!} k^{n-m}, \quad \text{if } n < N. \quad (A3b)$$

We then consider a function $f(k)$ whose asymptotic expansion ($k \rightarrow \pm\infty$) is of the following type

$$f(k) = k^{-p} \sum_{n=1-p}^N f_n k^{-2n} + O(k^{-p-2N-2}), \quad p = 0 \text{ or } 1. \quad (A4)$$

We also assume that $f(k)$ and all its derivatives are finite on the real axis, and that the asymptotic behaviors of the derivatives of $f(k)$ may be obtained differentiating Eq (A4).

For functions of this type we write the integral identity

$$\int_{-\infty}^{+\infty} dk k^{2n+p} f(k) - k^{-p} \sum_{m=1-p}^n f_m k^{-2m} = (-)^p \frac{1}{(2n+p)!} \int_{-\infty}^{+\infty} dk D^{2n+p} f(k). \quad (A5)$$

The validity of this relation is easily proved by partial integration and using Eqs. (A3a) and (A4). Of course the value of p in Eqs. (A4) and (A5) must be the same.

Another useful formula, whose proof requires only one extra partial integration, is the following:

$$\int_{-\infty}^{+\infty} dk k^{2n+p+1} \frac{d}{dk} \left[f(k) - k^{-p} \sum_{m=1-p}^n f_m k^{-2m} \right] = (-)^{p+1} \frac{(2n+p+1)}{(2n+p)!} \int_{-\infty}^{+\infty} dk D^{2n+p} f(k) \quad (A6)$$

Again the value of p must be the same as in Eq. (A4).

It is easily seen that Eqs. (A5) and (A6) remain valid if the integral runs only over one semiaxis, if the function $f(k)$ has a definite parity, so that the integrands in Eqs. (A5) and (A6) are even.

Finally we write integral expressions for the asymptotic coefficients f_n themselves:

$$f_n = (-)^{p+1} \frac{1}{(2n+p-1)!} \int_0^{\infty} dk \frac{1}{k} D^{2n+p} f(k) \quad (A7)$$

These are easily proved by partial integration and using Eq. (A3a). Again the value of p must be the same as in Eq. (A4). Note that the integrand in the right-hand side of this equation is an odd function of k , if $f(k)$ has the same parity of its asymptotic expansion.

APPENDIX B

In this appendix (that coincides with Appendix C of the paper by *Calogero and Degasperis* [1968]), we report for completeness an explicit elementary derivation of the well known formula, Eq. (72), that relates the derivative of the (*S*-wave) scattering phase shift $\eta(k)$ produced by the potential $V(r)$ to the increment $n_f(E)$ in the density of states due to the same potential.

We enclose the system in a spherical box of radius R . In the absence of potential the allowed states are characterized by the wave functions $\sin(k_n r)$, with

$$k_n R = n\pi \quad (B1)$$

Thus the density of states in this case is

$$n_0(E) = \frac{\Delta n}{\Delta E} = \frac{\Delta n}{2k \Delta k} = \frac{R}{2\pi k} \quad (B2)$$

On the other hand, if the potential is present and if R is large enough so that the asymptotic expression $\sin[kR + \eta(k)]$ may be used in place of the radial wave function, then in place of Eq. (B1) we have

$$k_n R + \eta(k_n) = n\pi \quad (B3)$$

Here of course $\eta(k)$ is the phase shift due to the potential. Thus the density of states in this case is

$$n(E) = \frac{\Delta n}{\Delta E} = \frac{1}{2\pi k} R + \frac{\Delta \eta}{\Delta k} \quad (B4)$$

But by definition

$$n_f(E) = n(E) - n_0(E) \quad (B5)$$

and inserting in this equation the expressions (B2) and (B4) we get, in the limit of the continuum ($R \rightarrow \infty$, $\Delta k \rightarrow dk$), Eq. (72), *q.e.d.*

APPENDIX C

This appendix coincides with appendix E of the paper by Calogero and Degasperis [1968]. In it we investigate the expression

$$\frac{1}{\pi} \int_{-\infty}^{+\infty} dr \int_{-\infty}^{+\infty} dk \exp(ikr) [(T+V)^D - T^D] \exp(ikr) \quad (C1)$$

where

$$T = - \frac{d^2}{dr^2} \quad (C2)$$

and V is a holomorphic function of r which vanishes sufficiently fast at infinity to allow any number of integrations by parts with neglect of the contributions at both extremes.

We begin evaluating the integral

$$\frac{1}{\pi} \int_{-\infty}^{+\infty} dr \int_{-\infty}^{+\infty} dk \exp(ikr) f(r) T g(r) \exp(ikr) = \frac{1}{\pi} \int_{-\infty}^{+\infty} dr \int_{-\infty}^{+\infty} dk \exp(2ikr) f(r) [-g''(r) - 2ikg'(r) + k^2 g(r)] \quad (C3)$$

where $f(r)$ and $g(r)$ are arbitrary functions which vanish asymptotically together with their derivatives. We use the identities

$$2ik \exp(2ikr) = \frac{d}{dr} \exp(2ikr) \quad (C4a)$$

and

$$k^2 \exp(2ikr) = - \frac{1}{4} \frac{d^2}{dr^2} \exp(2ikr) \quad (C4b)$$

to eliminate the terms linear and quadratic in k in the right-hand side. We then integrate by parts to eliminate the derivatives acting on the exponential and we thus secure

$$\begin{aligned} \frac{1}{\pi} \int_{-\infty}^{+\infty} dr \int_{-\infty}^{+\infty} dk \exp(ikr) f(r) T g(r) \exp(ikr) \\ = \frac{1}{4\pi} \int_{-\infty}^{+\infty} dr \int_{-\infty}^{+\infty} dk \exp(2ikr) [2f'(r)g'(r) - f''(r)g(r) - f(r)g''(r)] \end{aligned} \quad (C5)$$

Finally using the expression

$$\delta(r) = \frac{1}{\pi} \int_{-\infty}^{+\infty} dk \exp(2ikr) \quad (C6)$$

where δ is the Dirac distribution, we obtain

$$\frac{1}{\pi} \int_{-\infty}^{+\infty} dr \int_{-\infty}^{+\infty} dk \exp(ikr) f(r) Tg(r) \exp(ikr) = \frac{1}{4} \left[2f'(r)g'(r) - f''(r)g(r) - f(r)g''(r) \right] \Big|_{r=0} \quad (C7)$$

It is convenient at this point to introduce the two differential operators \vec{d} and \overleftarrow{d} , defined as follows:

$$f(r) \vec{d} g(r) = f(r) g'(r) \quad (C8a)$$

$$f(r) \overleftarrow{d} g(r) = f'(r) g(r) \quad (C8b)$$

We may then rewrite Eq. (C7) in the form

$$\frac{1}{\pi} \int_{-\infty}^{+\infty} dr \int_{-\infty}^{+\infty} dk \exp(ikr) f(r) Tg(r) \exp(ikr) = -\frac{1}{4} f(r) (\vec{d} - \overleftarrow{d})^2 g(r) \Big|_{r=0} \quad (C9)$$

On the other hand we also have, for an arbitrary function $f(r)$

$$\frac{1}{\pi} \int_{-\infty}^{+\infty} dr \int_{-\infty}^{+\infty} dk \exp(ikr) f(r) \exp(ikr) = f(0) \quad (C10)$$

From this equation and Eq. (C9) we infer that

$$\frac{1}{\pi} \int_{-\infty}^{+\infty} dr \int_{-\infty}^{+\infty} dk \exp(ikr) \left[(T+V)^P - T^P \right] \exp(ikr) = \left\{ 1 \cdot \left[V(r) - \frac{1}{4} (\vec{d} - \overleftarrow{d})^2 \right]^P \cdot 1 \right\} \Big|_{r=0} \quad (C11)$$

It is interesting to note a remarkable property of the operator $\vec{d} + \overleftarrow{d}$, namely

$$f(r) (\vec{d} + \overleftarrow{d}) g(r) = \frac{d}{dr} [f(r) g(r)] \quad (C12)$$

Thus the operator $\vec{d} + \overleftarrow{d}$ commutes with a function of r , namely its position within a product is irrelevant:

$$(\vec{d} + \overleftarrow{d}) f(r) g(r) = f(r) (\vec{d} + \overleftarrow{d}) g(r) = f(r) g(r) (\vec{d} + \overleftarrow{d}) \quad (C13)$$

This is, unfortunately, not the case for the operator $\vec{d} - \overleftarrow{d}$.

We may use this property of the operator $\vec{d} + \overleftarrow{d}$ to express the right-hand side of Eq. (C11) in the form

$$\left\{ 1 \cdot \left[V(r) - \frac{1}{4} (\vec{d} - \overleftarrow{d})^2 \right]^p \cdot 1 \right\} = \sum_{\ell=0}^{p-1} \binom{p}{\ell} (-1)^\ell 2^{-2\ell} \frac{d^{2\ell}}{dr^{2\ell}} \left\{ 1 \cdot [V(r) + \vec{d} \cdot \overleftarrow{d}]^{p-\ell} \cdot 1 \right\} \quad (C14)$$

which is obtained using the identity

$$(\vec{d} - \overleftarrow{d})^2 = (\vec{d} + \overleftarrow{d})^2 - 4\vec{d} \cdot \overleftarrow{d} \quad (C15)$$

and the binomial expansion, which is applicable thanks to the commutativity of $\vec{d} + \overleftarrow{d}$. This expansion may be useful because the operator $\vec{d} \cdot \overleftarrow{d}$ is such that

$$1 \cdot [V(r) + \vec{d} \cdot \overleftarrow{d}]^p \cdot 1 = V(r) [V(r) + \vec{d} \cdot \overleftarrow{d}]^{p-1} V(r), \quad p = 2, 3, \dots \quad (C16)$$

moreover this term contains derivatives only up to the order $p-2$. On the other hand the quantity in the left-hand side of Eq. (C14) contains derivatives up to the order $2p-2$. Note that from Eqs. (C14) and (C16) one easily gets

$$1 \cdot \left[V(r) - \frac{1}{4} (\vec{d} - \overleftarrow{d})^2 \right]^p \cdot 1 = (-)^{p+1} p 2^{2-2p} \frac{d^{2p-2}}{dr^{2p-2}} V(r) \quad (C17)$$

+ terms containing derivatives of order $2p-4$ or less

APPENDIX D

In this appendix we derive explicit expressions for the coefficients g_n of the asymptotic expansion of $g(k)$, in terms of the (*S*-wave) scattering phase shift $\eta(k)$. The treatment given here is a simplified version of appendix G of the paper by *Calogero and Degasperis* [1968], the simplification resulting mainly from the assumption that no bound states are present. We recall that this assumption implies that the Jost function $f(k)$, and the function $g(k)$ of Eq. (24), are both holomorphic in the lower half of the complex k -plane.

We consider separately the expansions of the real and imaginary parts of $g(k)$, since Eq. (D17) obviously implies, for real k ,

$$\operatorname{Re} g(k) = \sum_{n=0}^N g_{2n+1} (-)^{n+1} (2k)^{-2n-2} + O(k^{-2N-4}), \quad (\text{D1})$$

$$\operatorname{Im} g(k) = \sum_{n=0}^N g_{2n} (-)^n (2k)^{-2n-1} + O(k^{-2N-3}). \quad (\text{D2})$$

We treat first the imaginary part. Let us recall that the explicit expression of this quantity for real k is

$$\operatorname{Im} g(k) = k \{1 - \exp[-2\Delta(k)]\} \quad (\text{D3})$$

where

$$\Delta(k) = -\frac{2}{\pi} P \int_0^\infty dq \frac{q}{q^2 - k^2} \eta(q) \quad (\text{D4})$$

It is convenient to discuss first the asymptotic expansion of $\Delta(k)$, that was introduced earlier:

$$\Delta(k) = \sum_{n=1}^N d_n k^{-2n} + O(k^{-2N-2}) \quad (\text{D5})$$

We now prove that

$$d_p = -\left[\frac{2}{\pi(2p-1)!} \right] \int_0^\infty dk D^{2p-1} \eta(k) \quad (\text{D6})$$

This formula has been quoted from the paper of *Buslaev and Faddeev* [1960]. The operator D is that defined in Appendix A.

Before providing a proof of this formula we show how it may be obtained by a less rigorous but more straightforward approach. In fact if we substitute in Eq. (D4) the asymptotic expansion

$$(q^2 - k^2)^{-1} = - \sum_{p=1}^N q^{2p-2} k^{-2p} + O(k^{-2N-2}) \quad (D7)$$

we get formally

$$\Delta(k) = \sum_{p=1}^N k^{-2p} \frac{1}{\pi} \int_{-\infty}^{+\infty} dq q^{2p-1} \eta(q) + O(k^{-2N-2}) \quad (D8)$$

which implies Eq. (D6) because, by partial integration (see Appendix A)

$$- \left[\frac{1}{(2p-1)!} \right] \int_0^{\infty} dk D^{2p-1} \eta(k) = \int_0^{\infty} dk k^{2p-1} \eta(k) \quad (D9)$$

Of course, the last two equations need not make sense, because the integrals in the right-hand side generally do not converge. Both equations, however, become valid, provided an appropriate part of its asymptotic expansion is subtracted from the function $\eta(k)$, so as to make the integrals convergent. Specifically in place of $\eta(k)$ we must substitute, in both equations,

$$\eta(k) - \sum_{n=0}^{p-1} a_n k^{-2n-1} \quad (D10)$$

where the coefficients a_n are those of Eq. (33). It has been proved in Appendix A that with this substitution, Eq. (D9) holds true. There remains now to show that with this substitution also Eq. (D8) becomes correct.

To prove this result we introduce, after *Roberts* [1964b], the function

$$h(k) = \frac{d}{dk} \ln f(k) \equiv \frac{df(k)/dk}{f(k)} \quad (D11a)$$

$$= \frac{d\Delta(k)}{dk} + i \left[\frac{d\eta(k)}{dk} \right] \quad (D11b)$$

and its asymptotic expansion

$$h(k) = i \sum_{n=0}^N (-2ik)^{-n-2} h_n + O(k^{-N-3}) \quad (D12)$$

Equation (D11b) follows from Eqs. (21 - 22). A comparison of this expansion with those for $\Delta(k)$ and $\eta(k)$, Eqs. (D5) and (33), implies through Eq. (D11b) the relations

$$d_p = (-)^p \frac{1}{p} 2^{-2p-2} h_{2p-1} \quad (D13)$$

$$a_p = (-)^p \left(\frac{1}{2p+1} \right) 2^{-2p-2} h_{2p} \quad (D14)$$

On the other hand, integrating the function

$$F(k) = k^{2p} \left[h(k) - i \sum_{n=0}^{2p-1} h_n (-2ik)^{-n-2} \right] \quad (D15)$$

along a contour composed of a large semicircle in the lower half k -plane and of the real axis indented at $k=0$, and using the fact that the function $h(k)$ is holomorphic in the lower half plane (because $f(k)$ is holomorphic and it has no zeros), one finds

$$(-)^p 2^{-2p-2} h_{2p-1} = -\frac{1}{2\pi} \int_{-\infty}^{+\infty} dk k^{2p} \frac{d}{dk} \left[\eta(k) - \sum_{n=0}^{p-1} a_n k^{-2n-1} \right] \quad (D16)$$

From this equation and Eq. (D13) we get

$$d_p = -\frac{1}{2\pi p} \int_{-\infty}^{+\infty} dk k^{2p} \frac{d}{dk} \left[\eta(k) - \sum_{n=0}^{p-1} a_n k^{-2n-1} \right] \quad (D17)$$

This equation, by Eq. (A6), becomes Eq. (D6).

There remains now to express the quantity g_{2p} in terms of the coefficients d_n . To achieve this aim we must first solve the following algebraic problem: let

$$z(x) = \exp[y(x)] \quad (D18)$$

$$y(x) = \sum_{n=1}^{\infty} y_n x^n \quad (D19)$$

$$z(x) = 1 + \sum_{n=1}^{\infty} z_n x^n \quad (D20)$$

and express z_p in terms of y_n .

The solution of this problem can be found noting that

$$z_n = \frac{z^{(n)}(0)}{n!} \quad (D21)$$

and

$$y_n = \frac{y^{(n)}(0)}{n!} \quad (D22)$$

the parenthetical superscript indicating the order of differentiation, and using the formula for the differentiation of a composite function, namely,

$$\frac{d^n}{dx^n} F[y(x)] = n! \sum \frac{d^m}{dy^m} F[y(x)] \prod_{j=1}^{\ell} \left\{ \frac{[y^{(j)}(x)/j!]^{s_j}}{s_j!} \right\} \quad (D23)$$

where

$$m = \sum_{j=1}^{\ell} s_j \quad (D24)$$

and the sum is extended over all positive integral values of ℓ and s_ℓ and nonnegative integral values of s_j , $j = 1, 2, \dots, \ell - 1$, such that

$$\sum_{j=1}^{\ell} j s_j = n \quad (D25)$$

We find in this manner

$$z_p = \sum_{\ell=1}^{\infty} \prod_{j=1}^{\ell} \frac{(y_j)^{s_j}}{s_j!} \quad (\text{D26})$$

this sum being now extended over all the positive integral values of ℓ and s_ℓ and the nonnegative values of s_j , $j = 1, 2, \dots, \ell - 1$ such that

$$\sum_{j=1}^{\ell} j s_j = p \quad (\text{D27})$$

But since

$$\text{Im } g(k) = k \{1 - \exp[-2\Delta(k)]\} \quad (\text{D28})$$

we may now set

$$x = k^{-2} \quad (\text{D29})$$

$$z(x) = 1 - k^{-1} \text{Im } g(k) \quad (\text{D30})$$

$$y(x) = -2\Delta(k) \quad (\text{D31})$$

Then from Eqs. (D2) and (D20) we find

$$g_{2p} = (-)^{p+1} 2^{2p+1} z_{p+1} \quad (\text{D32})$$

and from Eqs. (D5) and (D19)

$$y_n = -2 d_n \quad (\text{D33})$$

Thus in conclusion

$$g_{2p} = (-)^{p+1} 2^{2p+1} \sum_{\ell=1}^{\infty} \prod_{j=1}^{\ell} \frac{(-2d_j)^{s_j}}{s_j!} \quad (\text{D34})$$

the sum being extended over all positive integral values of ℓ and s_ℓ and nonnegative integral values of s_j , $j = 1, 2, \dots, \ell - 1$ such that

$$\sum_{j=1}^{\ell} j s_j = p + 1 \quad (\text{D35})$$

Note that this last equation implies that the maximum value of ℓ in Eq. (D34) cannot exceed

$$\ell_{\max} = p + 1 \quad (\text{D36})$$

Let us proceed now to discuss the odd coefficients g_{2p+1} , namely those appearing in the asymptotic expansion of $\text{Re } g(k)$ for real k . We use again the technique which we used above, except for the fact that we now integrate the function

$$\bar{F}(k) = k^{2p+1} \left[g(k) - \sum_{n=0}^{2p+1} g_n (-2ik)^{-n-1} \right] \quad (\text{D37})$$

The contour is the same as in the previous case; as $\bar{F}(k)$ is again holomorphic in the lower half plane. Considering only the imaginary part of the result we secure in this manner

$$g_{2p+1} = (-)^{p+1} \frac{2^{2p+2}}{\pi} \int_{-\infty}^{+\infty} dk k^{2p+1} \left[\text{Im } g(k) - \sum_{n=0}^p (-)^n g_{2n} (2k)^{-2n-1} \right] \quad (\text{D38})$$

which, using Eqs. (25) and (A5), can be written in the form

$$g_{2p+1} = (-)^{p+1} \left[\frac{2^{2p+2}}{\pi(2p+2)!} \right] \int_{-\infty}^{+\infty} dk D^{2p+2} [1 - |f(k)|^{-2}] \quad (\text{D39a})$$

$$= (-)^{p+1} \left[\frac{2^{2p+2}}{\pi(2p+3)!} \right] \int_{-\infty}^{+\infty} dk D^{2p+3} [k^{-1} |f(k)|^{-2}] \quad (\text{D39b})$$

The second line of this equation is obtained from the first by an integration by parts.

Of course this equation might have been inferred, in a straightforward but nonrigorous way, expanding directly under the integral sign in Eq. (28), as was done at the beginning of this Appendix to obtain the coefficients d_p .

DISCUSSION

Unidentified speaker: Is there a way of relating more directly these coefficients in the asymptotics of η to the potential?

Calogero: Yes. This is a question concerning the direct problem. There are explicit formulas for that.

Grossi: Would it be possible for a monochromatic particle to express some of the functions that lead to the V_0 and its derivative as a function of the input parameter of the particle.

Calogero: This is the other problem, the problem at fixed energy and varying angular momenta. I have not looked into this.

THE INVERSE SCATTERING PROBLEM AT FIXED ANGULAR MOMENTUM FOR NONLOCAL SEPARABLE INTERACTIONS

K. Chadan

Laboratoire de Physique Théorique et Particules Élémentaires

Complexe Scientifique d'Orsay, Université de Paris - Sud

ABSTRACT

N73-11611

As is the major thrust of Newton's paper, the problem of inverse scattering at fixed angular momentum is considered. The problem is particularized to the case of nonlocal separable interactions.

This paper is a brief survey of the inverse problem for nonlocal separable interactions. As we shall see, this problem can be solved exactly by integration. In fact, it amounts to solving singular integral equations of the Hilbert-Mushkhelishvili type, which have been studied extensively in the past and appear in many areas of physics, including theory of elasticity and dispersion relations in high energy physics.

INTRODUCTION

The Schrödinger equation for the scattering of a nonrelativistic particle of mass M by a general nonlocal interaction reads:

$$(\Delta_r + E)\psi(k, r) = \int U(r, r')\psi(k, r')dr' \quad (1)$$

where we have chosen the units such that $\hbar = 2M = 1$, $E = \frac{1}{2}k^2/2M = k^2$, k being the wave number. We assume now that

$$U(r, r') = \sum_{\ell} \epsilon_{\ell} u_{\ell}(r)u_{\ell}(r')P_{\ell}(\cos \theta) \quad (2)$$

$$r = |r| \quad r' = |r'| \quad \cos \theta = \frac{r \cdot r'}{rr'} \quad \epsilon_{\ell} = \pm 1$$

We shall study later separable interactions of a more general kind. Making the usual partial wave decomposition

$$\psi(k, r) \sim \sum (2\ell + 1)i^{\ell} \frac{\psi_{\ell}(k, r)}{kr} P_{\ell}[\cos(k, r)] \quad (3)$$

we obtain the reduced radial equation

$$\left\{ \frac{d^2}{dr^2} + \left[E - \frac{\ell(\ell+1)}{r^2} \right] \right\} \psi_\ell(k, r) = \epsilon_\ell U_\ell(r) \int_0^\infty \psi_\ell(k, r') U_\ell(r') dr' \quad (4)$$

$$U_\ell(r) = \sqrt{4\pi} r u_\ell(r)$$

to which we have to add the boundary condition $\psi_\ell(k, 0) = 0$. The asymptotic form of ψ_ℓ at large distances should be

$$\psi_\ell(r \rightarrow \infty) \sim A_\ell \sin \left[kr - \frac{1}{2} \ell\pi + \delta_\ell(k) \right] \quad (5)$$

where $\delta_\ell(k)$ is the phase shift.

In each angular momentum state — that is, for each ℓ — the inverse problem consists of finding the “potential” $U_\ell(r)$ from the knowledge of $\delta_\ell(k)$, which is assumed to be known for all positive energies, and the energies of the eventual bound states (i.e., square-integrable solutions of Eq. (4)). These later are divided into two groups: those with negative energy, and those with positive energy.

For simplicity, we shall consider from now on the case of S-wave ($\ell = 0$). All the results can immediately be generalized to the case of higher waves, as we shall see. Let us therefore look first at scattering solutions.

SCATTERING SOLUTIONS

We have to solve

$$\psi'' + E\psi = \epsilon N(k)U(r) \quad E = k^2 \quad \epsilon = \epsilon_0 \quad (6)$$

$$N(k) = \int_0^\infty U(r') \psi(k, r') dr'$$

This equation can easily be solved by taking the Fourier sine transform*

$$\tilde{\psi}(k, p) = \int_0^\infty \psi(k, r) \frac{\sin pr}{p} dr \quad (7)$$

$$\psi(k, r) = \frac{2}{\pi} \int_0^\infty \tilde{\psi}(k, p) \frac{\sin pr}{p} p^2 dp \quad (8)$$

*The Fourier transform of the wave-functions is not a function but a distribution. However, there is no problem in handling such harmless objects. We use only Fourier sine transform since $\psi(k, 0) = 0$.

One then immediately finds from the asymptotic form of ψ that the phase shift is given by

$$-\frac{2k}{\pi} \operatorname{tg} \delta(k) = \frac{F(k)}{1 + P \int_0^\infty \frac{F(p)}{p^2 - k^2} dp} \quad (9)$$

$$F(k) = \frac{2}{\pi} \epsilon k^2 [\tilde{U}(k)]^2 \quad (10)$$

where P denotes the principal value of the integral, and \tilde{U} is the Fourier sine transform of the potential:

$$\tilde{U}(k) = \int_0^\infty U(r) \frac{\sin kr}{k} dr \quad (11)$$

As it is obvious now, we have to impose some integrability conditions on the potential in order for the above formulas to be valid. It is easily seen that the following rather general conditions are sufficient.

U is locally integrable, except perhaps at the origin,

$$\int_0^\infty r^{1/2-\eta} |U(r)| dr < \infty \quad (12)$$

$$\int_0^\infty r^{1+\eta'} |U(r)| dr < \infty \quad (13)$$

where η and η' are as small as we wish, but positive. (We use these symbols to denote small positive quantities. Their actual values may differ from place to place.)

Condition (12) ensures that the boundary condition is satisfied and that the integral appearing in the right-hand side of the radial Schrödinger equation is convergent at its lower extremity, whereas conditions (12) and (13) together guarantee the existence of this integral as a whole, the existence of $\tilde{U}(k)$ for all energies ($k \geq 0$), and the convergence at infinity of the integral

$$G(k^2) = P \int_0^\infty \frac{F(p)}{p^2 - k^2} dp \quad (14)$$

which appears in the denominator of Eq. (9).

We notice now that since $\tilde{U}(r) \in L_1[0, \infty]$, $k\tilde{U}(k)$ is everywhere continuously differentiable. (One can differentiate Eq. (11) under the integral sign.) Therefore, the principal value integral (14) is meaningful and defines an

everywhere Hölder-continuous function with Hölder index $1 - \epsilon$, ϵ as small as we wish; in fact, one knows that $|G(E+h) - G(E)| = O(|h| \log |h|)$. Even more can be said about the integral (14). It can be shown that $F(k)$, for $k \rightarrow \infty$, behaves like $k^{1-\eta}$, where η is some positive quantity. It then follows that the principal value integral vanishes when $k \rightarrow \infty$. As a consequence, we find, according to Eqs. (9) and (10) that $\text{tg } \delta(\infty) = 0$; and for k large enough, the sign of $\text{tg } \delta(k)$ is that of $-\epsilon$.

BOUND STATES

We now look at the square-integrable solutions of the radial equation. First, it can be shown that the bound states with positive energy are characterized by the simultaneous roots of

$$F(k) = 0 \quad (15)$$

$$1 + P \int_0^\infty \frac{F(p)}{(p^2 - k^2)} dp = 0 \quad (16)$$

and that their number is finite. Equation (16) is the eigenvalue equation, which also gives the energy of the true bound state (the bound state with negative energy), whereas Eq. (15) ensures that the radial equation has one solution that vanishes in the limit of $r \rightarrow \infty$. For such energies, there is also the scattering solution, and these two solutions are independent of each other. Moreover, it can easily be seen that at the energies of these bound states the phase shift goes through $n\pi$, with n an integer ($n \geq 0$). Indeed, it is clear from the definition (10) that the roots of Eq. (15) are of even multiplicity (at least of order 2). Therefore, if k_0 is such a root, we have

$$\frac{\partial}{\partial k} \left[1 + P \int_0^\infty \frac{F(p)}{(p^2 - k^2)} dp \right] \Big|_{k=k_0} = 2k_0 \int_0^\infty \frac{F(p)}{(p^2 - k_0^2)^2} dp \neq 0 \quad (17)$$

where we have written $F(p) = F(p) - F(k_0)$. As is obvious from (10), this quantity has the sign of ϵ . The last integral is meaningful since $F(p)$ has a zero of order at least two at $p = k_0$. It follows that if the roots of Eqs. (15) and (16) coincide the root of the latter is simple. From the definition (9) it follows immediately that at these energies the phase shift crosses the value $n\pi$. Moreover, it does so downward as is seen by differentiating Eq. (9), and using the fact that $\epsilon(17) > 0$. Once the phase shift has crossed a value $n\pi$, it can never cross it again. Of course, it can reach again the same value $n\pi$ at some other energy, but this happens when only the numerator of Eq. (9) vanishes, and so δ , which has a zero slope at this energy, does not cross the value $n\pi$.

As for the bound state with negative energy, it is given by the root of the eigenvalue equation ($E = -x$)

$$H(x) = 1 + \int_0^\infty \frac{F(p)}{p^2 + x} dp = 0 \quad x > 0 \quad (18)$$

It is now clear that there is no true bound state if $\epsilon = 1$, since in that case $F(p)$ is positive. We therefore may call such separable potentials repulsive. However, when $\epsilon = -1$, and

$$H(0) = 1 + \int_0^{\infty} \left(\frac{F(p)}{p^2} \right) dp < 0 \quad (19)$$

it is obvious that Eq. (18) has one positive root x_0 , corresponding to a bound state of energy $-x_0$.

Figure 1 summarizes typical behaviors of the phase shift. Because of $\lim_{k \rightarrow \infty} \delta(k) = 0$, we can always choose δ such that

$$\delta(\infty) = 0 \quad (20)$$

We adhere to this determination from now on. Therefore, the Levinson theorem, which relates $\delta(0) - \delta(\infty)$ to the total number of bound states, reads

$$\delta(0) = N\pi \quad (21)$$

$$N = \nu + n \quad (22)$$

where ν is the number of bound states of positive energy, and n is the number of true bound states. We leave out the possibility of a resonance at zero energy. If this happens we must add $\pi/2$ to the r.h.s. of Eq. (21). Since the true bound state occurs only when the sign of the potential is $\epsilon = -1$, and since, as we saw before, the sign of $\delta(k)$ for sufficiently large k is opposite to that of ϵ , it follows that if the interaction admits a true bound state, $\delta(k)$ must be positive (small) when k is large.

Therefore, just by inspecting the sign of the phase shift at large energies, we can determine ϵ .

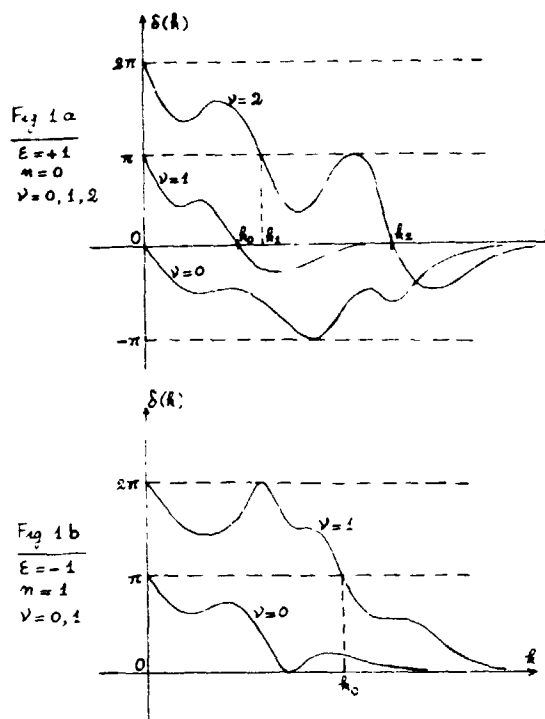


Figure 1a. and Figure 1b.

THE INVERSE PROBLEM

To determine $U(r)$ from the phase shift, we must first assume that the latter meets all the requirements for the problem to have a solution. For instance, if the phase shift crosses upward the value $n\pi$, n integer, when we increase k , we are sure that our problem with one separable potential has no solution.

We consider first the inverse problem for the case of $\epsilon = 1$. As we saw before, $\delta(k)$ at large energies must be small but negative. The eventual bound states with positive energy are then determined by just looking when δ crosses (upward!) the values $n\pi$, $n = 0, 1, 2, \dots$, as we decrease k from infinity. At $k = 0$, we must have $\delta(0) = \nu\pi$. To find the potential from the phase shift, we have to solve the integral equation (9) for $F(k)$. Once this quantity is found, there is no problem for finding $\tilde{U}(k)$ via Eq. (9), and then undoing the Fourier transform (11). For simplicity, we consider the case $\nu = 1$. We may write Eq. (9) in the form

$$\phi(x) = 1 + \frac{1}{\pi} \int_0^{\infty} \frac{h^*(y)\phi(y)}{y-x-i0} dy \quad (23)$$

where

$$x = k^2 \quad (24)$$

$$\tilde{\delta}(x) \equiv \delta(\sqrt{x}) \quad (25)$$

$$g(x) = -\frac{2\sqrt{x}}{\pi} \operatorname{tg} \tilde{\delta}(x) \quad (26)$$

$$\phi(x) = \frac{F(x)}{g(x)} \left[1 + i\pi \frac{g(x)}{2\sqrt{x}} \right] \quad (27)$$

$$h(x) = \frac{\pi g(x)}{2\sqrt{x}} \left[1 - i\pi \frac{g(x)}{2\sqrt{x}} \right]^{-1} = -\sin \tilde{\delta} e^{-i\tilde{\delta}} \quad (28)$$

We assume now, according to our previous analysis, that δ is Hölder-continuous with some positive index and that it behaves like $k^{-\eta}$, η some positive quantity, when $k \rightarrow \infty$. These conditions are necessary if we wish to find a potential satisfying Eqs. (12) and (13). They are also sufficient according to what follows.

To solve the integral equation (23), we consider first the function

$$H(z) = 1 + \frac{1}{\pi} \int_0^\infty \frac{h(y)\phi(y)}{y-z} dy \quad (29)$$

Assuming a priori the Hölder-continuity of ϕ and the proper convergence of the integral, we see that $H(z)$ is analytic in the z plane cut from 0 to $+\infty$ (fig. 2). Moreover,

$$\lim_{|z| \rightarrow \infty} H(z) = 1 \quad (30)$$

in all directions.

From the integral Eq. (23), we see that its solution is necessarily of the form

$$\phi(x) = H(x_+) \equiv \lim_{\epsilon \rightarrow +0} H(x + i\epsilon) \quad 0 \leq x < \infty \quad (31)$$

On the other hand, the discontinuity of $H(z)$ on the cut is given by

$$H(x_+) - H(x_-) = -2i \sin \tilde{\delta}(x) e^{i\tilde{\delta}(x)} \phi(x) \quad (32)$$

Upon substituting ϕ by its value (31) we find

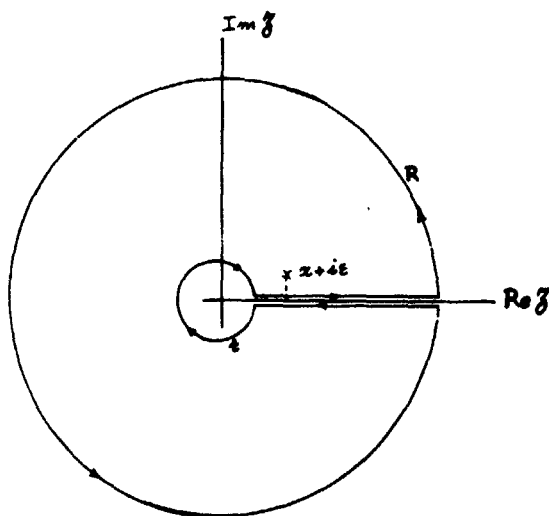


Figure 2.— The Contour Γ .

$$e^{2i\tilde{\delta}(x)} H(x_+) - H(x_-) = 0 \quad 0 \leq x < \infty \quad (33)$$

which is a homogeneous Riemann-Hilbert type equation. It is now easily verified that a particular solution of this equation, satisfying Eq. (30) is given by

$$H(z) = \exp [\omega(z)] \quad (34)$$

$$\omega(z) = -\frac{1}{\pi} \int_0^\infty \frac{\tilde{\delta}(y)}{y-z} dy \quad (35)$$

Indeed, from our assumptions on the phase shift, we have

$$\lim_{|z| \rightarrow \infty} \omega(z) = 0 \quad (36)$$

in all directions. Moreover, $\omega(z)$ is a well-defined function on the cut, except perhaps at $z = 0$, where its behavior is given by

$$\omega(z) = \frac{\tilde{\delta}(0)}{\pi} \log z + \tilde{\omega}(z) \quad (37)$$

$\tilde{\omega}(z)$ being finite when $z \rightarrow 0$, and $\tilde{\delta}(0) = \delta(0) = \pi$ (one positive energy-bound state). Therefore, the function $\exp[\omega(z)]$ has only a simple zero at $z = 0$. According to Eqs. (34) and (35), $\phi(x)$ is given by

$$\phi(x) = e^{\omega(x_+)} = e^{\alpha(x)} e^{-i\tilde{\delta}(x)} \quad (38)$$

$$\alpha(x) = -\frac{1}{\pi} P \int_0^\infty \frac{\tilde{\delta}(y)}{y-x} dy \quad (39)$$

We have to verify now that this indeed satisfies the integral equation (23). This is most easily done since by Cauchy's theorem we have

$$\begin{aligned} -\frac{1}{\pi} \int_0^\infty \frac{e^{i\tilde{\delta}(y)} \sin \tilde{\delta}(y)}{y-x-i0} e^{\alpha(y)} e^{-i\tilde{\delta}(y)} dy &= \lim_{\epsilon \rightarrow 0} \left[\frac{1}{2i\pi} \left(\int_\Gamma \frac{e^{\omega(z)}}{z-x-i\epsilon} dz - \int_R \dots \right) \right] \\ &= e^{\omega(x_+)} - 1 = \phi(x) - 1 \end{aligned} \quad (40)$$

where the contour Γ is shown on figure 2, R being the circle with a very large radius. (With $\omega(z)$ regular at $z = 0$, it is obvious that the contribution of the small circle enclosing the origin vanishes when its radius r tends to zero.)

Note that $\phi(x)$ is regular at $x = 0$. In fact it has a simple zero there. It is also obvious from Eqs. (38) and (39) that $\phi(x)$ is Hölder-continuous with the same index as the phase shift, and remains bounded when $x \rightarrow \infty$. Our a priori hypothesis on ϕ for solving the integral equation is therefore justified.

Let us consider now the homogeneous equation

$$\phi_0(x) = \frac{1}{\pi} \int_0^{\infty} \frac{h^*(y)}{y-x-i0} \phi_0(y) dy \quad (41)$$

to which we associate the function

$$H_0(y) = \frac{1}{\pi} \int_0^{\infty} \frac{h^*(y)}{y-z} \phi_0(y) dy \quad (42)$$

Assuming a priori that ϕ_0 is Hölder-continuous ... etc, we get

$$\lim_{|z| \rightarrow \infty} H_0(z) = 0 \quad (43)$$

in all directions. Proceeding as for H , we obtain once more

$$e^{2i\tilde{\delta}(x)} H_0(x_+) - H_0(x_-) = 0 \quad (44)$$

Using the Ansatz

$$H_0(z) = \sum_s A_s \frac{e^{\omega(z)}}{z^s} \quad (45)$$

where $\omega(z)$ is again defined by Eq. (35) and A_s are arbitrary constants, we find

$$\phi_0(x) = \frac{A}{x} e^{\omega(x_+)} \quad (46)$$

Again it can be checked by contour integration that Eq. (46) is indeed a solution of Eq. (41), and that it has all the required properties.

The general solution of our integral equation is therefore

$$\phi(x) = \left(1 + \frac{A}{x}\right) e^{\omega(x_+)} \quad (47)$$

Going back to our earlier notations, we obtain

$$F(k) = -\frac{2k}{\pi} \sin \delta(k) e^{\alpha(k^2)} \left(1 + \frac{A}{k^2}\right) \quad (48)$$

$$\alpha(k^2) = -\frac{2}{\pi} P \int_0^\infty \frac{\delta(p)}{p^2 - k^2} p \, dp \quad (49)$$

This solution depends on the arbitrary constant A . However, from its definition, F must keep a constant sign for all values of k . Here F should be positive since $\epsilon = 1$. On the other hand, we know that $\sin \delta$ changes sign at the energy of the bound state, which occurs at $E = k_0^2$, k_0 being the point where the phase shift crosses the value zero [fig. 1(a)]. It follows that we must have

$$A = -k_0^2 \quad (50)$$

in order for F given by Eq. (48) to keep a constant (positive) sign. We therefore see that the solution is completely determined by the phase shift alone, since this also determines k_0 . Once F is known, we obtain

$$k \tilde{U}(k) = \sqrt{F(k)} \quad (51)$$

which gives in turn $U(r)$ by inverting the Fourier transform (11). It can then be checked that U satisfies the conditions (12) and (13) since F is Hölder-continuous and decreases like $k^{-\eta}$, $\eta > 0$, when $k \rightarrow \infty$. From Eqs. (48) and (49) it is clear that these properties hold for F if they hold for the phase shift.

The case when several positive energy bound states occur at $E_\alpha = k_\alpha^2$, $\alpha = 1, 2, \dots, \nu$, and $\epsilon = 1$, is quite similar, and leads to

$$F(k) = -\frac{2k}{\pi} \sin \delta(k) e^{\alpha(k^2)} \prod_\alpha \left(1 - \frac{k_\alpha^2}{k^2}\right) \quad (52)$$

We consider now the case where $\epsilon = -1$, and where there are two bound states, one with positive energy $E_0 = k_0^2$, and the other a true bound state with negative energy $E = -\gamma^2$. A typical behavior of the phase shift in this case is shown on Figure 1(b). Since $\delta(0) = 2\pi$, we get that $\omega(z)$ defined by Eq. (35) has, according to Eq. (37) a double zero at the origin. Proceeding as before, and taking into account the eigenvalue equation (18) with $x = \gamma^2$, and the fact that the solution $F(k)$ must keep a constant (negative) sign for all k , we obtain

$$F(k) = -\frac{2k}{\pi} \sin \delta(k) e^{\alpha(k^2)} \left(1 - \frac{k_0^2}{k^2}\right) \left(1 + \frac{\gamma^2}{k^2}\right) \quad (53)$$

In this case, we see that the potential is determined unambiguously from the phase shift and the energy of the true bound state.

The generalization of the above method to higher partial waves is now obvious. We only have to use the Hankel integral transform

$$\tilde{\psi}(k, p) = \int_0^{\infty} \psi(k, r) \frac{\sqrt{(kr)} J_{l+1/2}(kr)}{k^{l+1}} dr \quad (54)$$

instead of the sine transform (11). This permits us to eliminate the centrifugal potential and to obtain equations quite similar to those obtained for the S wave. The conclusions are then quite similar.

MORE GENERAL INTERACTIONS

The method we have been discussing can be used with more general nonlocal interactions. The first generalization consists in considering the interaction

$$U(r, r') = V(r) \delta(r - r') + \sum_l \epsilon_l U_l(r) u_l(r') P_l(\cos \theta) \quad (55)$$

where the local potential V is assumed to be known. The radial equation now reads

$$\psi_l'' + \left[E - \frac{l(l+1)}{r^2} \right] \psi_l = V \psi_l + \epsilon_l U_l(r) \int_0^{\infty} U_l(r') \psi_l(k, r') dr' \quad (56)$$

Here, the inverse problem consists of finding U_l when the phase shift and the local potential V are both known. For this purpose, we assume that the local potential satisfies

$$\int_0^{\infty} r |V(r)| dr < \infty \quad (57)$$

We can now solve (in principle) the radial equation with V alone:

$$\phi_l'' + \left[E - \frac{l(l+1)}{r^2} \right] \phi_l = V(r) \phi_l \quad (58)$$

$$\phi_l(r) \underset{r \rightarrow 0}{=} \frac{r^{l+1}}{(2l+1)!!} [1 + o(1)] \quad (59)$$

and define the integral transforms

$$\tilde{U}_\ell(E) = \int_0^\infty U_\ell(r) \phi_\ell(E, r) dr \quad (60)$$

$$U_\ell(r) = \int_{-\infty}^\infty \tilde{U}_\ell(E) \phi_\ell(E, r) d\rho(E) \quad (61)$$

where $\rho(E)$ is an appropriate measure, given in terms of the Jost function of V . (These integral transforms can be shown to have properties quite similar to those of Hankel transforms used earlier. They reduce, of course, to the latter when $V=0$. For the definition of the spectral measure $\rho(E)$, see Newton's discussion at the beginning of this chapter.) If we write the phase shift as (for simplicity, dropping the subscript ℓ)

$$\delta = \delta_V + \delta_U^V \quad (62)$$

where δ_V is the known phase shift due to V alone — that is, the phase-shift of Eq. (58) — we obtain for δ_U^V an equation quite similar to (9) and (10) where \tilde{U} is now given by Eq. (60). Therefore, we again have a Hilbert-Mushkelishvili-type equation for the inverse problem, which can be solved under the same hypotheses, as before, on δ_U^V . The conclusions are quite similar: if we know δ , and therefore δ_U^V , and the energies of true bound states ($E < 0$), and if δ_U^V satisfies certain conditions similar to those discussed earlier, we can find $U(r)$ unambiguously. (Here, too, one can show that the positive energy bound states are given by those values of k where $\delta_U^V(k)$ crosses downward $n\pi$ when we increase k .)

Another generalization consists of having several separable potentials in each angular momentum state:

$$U(r, r') = V(r) \delta(r - r') + \sum_{\ell} \sum_{s=1}^{S_\ell} \epsilon_{\ell s} U_{\ell s}(r) \dots \quad (63)$$

The right-hand side of the radial equation then reads

$$V \psi_\ell + \sum_{s=1}^{S_\ell} \epsilon_{\ell s} U_{\ell s}(r) \int_0^\infty U_{\ell s}(r') \psi_\ell(r') dr' \quad (64)$$

There are two ways of solving this equation: either we solve it step by step, using the integral transformed defined by the wave-function of the previous step, exactly as for Eq. (56); or we use the integral transform defined by ϕ_ℓ throughout, leading to a set of algebraic equations, to determine the relation between the potentials $U_{\ell s}$ and the phase shift. Once more, one is led to singular integral equations of the Hilbert-Mushkelishvili type, and finally to conclusions quite similar to those reached before the inverse problem. In this case, whatever δ_ℓ and the negative energy bound states may be, we can always find an appropriate set of $U_{\ell s}$ that solves the problem. In other words, δ_ℓ may have any shape if we choose S_ℓ large enough.

There has been recently some revival of interest in the use of separable interactions for studying the properties of the nuclear matter. For references, see Sabatier's paper earlier in this chapter.

We end up by giving as an amusing and somewhat surprising example the construction of a class of completely transparent nonlocal potentials. We start from any totally repulsive (everywhere positive) well-behaved local potential, that satisfies Eq. (57). We can then write its phase shift for any given ℓ , $\delta_V^{(\ell)}$, in the form

$$\delta_V^{(\ell)} = - \sum_{s=1}^{S_\ell} \delta^{(\ell s)} \quad (64)$$

where each $\delta^{(\ell s)}$ has the appropriate shape for the inverse problem for

$$\delta_V + \sum_{s=1}^{S_\ell} \delta^{(\ell s)} (\equiv 0)$$

to be solvable step by step as before. We end up in this way with a total phase shift that is zero at all energies. Doing this for every ℓ , we obtain the nonlocal potential (63) for which all the phase shifts are zero for all energies. Notice that there is no convergence problem for the sum over s in Eq. (63) since for ℓ large enough, the local phase shift $\delta_V^{(\ell)}$ is very small at all energies. This means that there is some ℓ beyond which we can take $S_\ell = 1$. The local potential and the decomposition of $\delta_V^{(\ell)}$ into the sum (65) being to a large extent arbitrary, we obtain thus a large class of nonlocal completely transparent potentials.

The same method can be used to obtain a class of nonlocal potentials, all corresponding to a given set of phase shifts, of which the above example is of course a particular case [$\delta_\ell(k) = 0$ for all ℓ and k].

These simple examples show that in general the inverse problem for nonlocal potentials has not necessarily a unique solution if one is only given the S matrix on the energy shell.

DISCUSSION

Degasperis: I have applied this theory, especially Calogero's work, to another field. Two years ago I applied a high-energy expansion to a plasma diagnostic problem. In that case, the high-frequency behavior of the electric potential outside a plasma cylinder provided the input data for reconstructing the electron density profile inside the plasma.

PARTICLE SCATTERING--CLASSICAL AND QUANTUM MECHANICAL

BIBLIOGRAPHY

- Agranovich, Z. S., and Marchenko, V. A., Reconstruction of the Potential Energy from the Dispersion Matrix, Usp. Mat. Nauk (N.S.), 12,1, 73, 143-144, 1957 a. (Am. Math. Soc. Transl., 16, 355).
- Agranovich, Z. S., and Marchenko, V. A., Re-establishment of the Potential from the Scattering Matrix for a System of Differential Equations, Dokl. Akad. Nauk SSSR, 113, 951-954, 1957 b.
- Agranovich, Z. S., and Marchenko, V. A., Construction of Tensor Forces from Scattering Data, Dokl. Akad. Nauk SSSR, 118, 1055-1058, 1958.
- Agranovich, Z. S., and Marchenko, V. A., The Inverse Problem of Scattering Theory, Gordon and Breach, New York, 1963.
- Alfaro, V. de, and Regge, T., Potential Scattering, Wiley, New York, 1965.
- Ambarzumian, V., Über eine Frage der Eigenwerttheorie, Z. Phys., 53, 690-695, 1929.
- Atkinson, D., and Calogero, F., Construction of the S-Matrix from its Left-Hand-Cut Discontinuity, When the Latter is Asymptotically Unbounded, Phys. Rev., 185, 1702-1716, 1969.
- Bargmann, V., Remarks on the Determination of a Central Field of Force from the Elastic Scattering Phase Shifts, Phys. Rev., 75, 301-303, 1949a.
- Bargmann, V., On the Connection Between Phase Shifts and Scattering Potential, Rev. Mod. Phys., 21, 488-493, 1949b.
- Benn, J., and Scharf, G., Determination of Nucleon-Nucleon Potentials From Scattering Data Using the Marchenko Theory, Nuclear Phys., A134, 481-504, 1969.
- Berezanskii, Yu. M., On the Uniqueness of the Determination of Schrödinger's Equation from its Spectral Function, Dokl. Akad. Nauk SSSR, 93, 591-594, 1953.
- Berezanskii, Yu. M., On the Inverse Problem of Spectral Analysis for the Schrödinger Equation, Dokl. Akad. Nauk SSSR, 105, 197-200, 1955.
- Berezanskii, Yu. M., The Uniqueness Theorem in the Inverse Problem of Spectral Analysis for the Schrödinger Equation, Tr. Mosk. Mater. Obsc., 7, 3-62, 1958 (Am. Math. Soc. Transl., 35, 167).
- Berezin, F. A., Spur Formula for the Schroedinger Many-Particle Equation, Dokl. Akad. Nauk SSSR, 157, 1069-1072, 1964 (English transl. Soviet Phys.-Doklady, 9, 641-644, 1965).
- Bethe, H. A., Theory of the Effective Range in Nuclear Scattering, Phys. Rev., 76, 38-50, 1949.
- Bhattacharjie, A., and Sudarshan, E. C. G., A Class of Solvable Potentials, Nuovo Cimento, 25, 864-879, 1962.
- Blatt, J. M., and Jackson, J. D., On the Interpretation of Neutron-Proton Scattering Data by the Schwinger Variational Method, Phys. Rev., 76, 18-36, 1949.

Blazek, M., The Gel'fand-Levitan Equation and its Application to the Scattering of Neutrons and Protons, Czech. J. Phys., B 12, 249-257, 1962a.

Blazek, M., Explicit Determination of a Potential with n Bound States by Means of the Solution of the Inverse Problem, Czech. J. Phys., B 12, 258-263, 1962b.

Blazek, M., On Some Equations for Potential Scattering, Czech. J. Phys., B 12, 497-507, 1962c.

Blazek, M., Determination of the Potential by the Analytic Properties of the Scattering Amplitude, Mat. Fys. Casopis SAV, 13, 147-175, 1963.

Blazek, M., On the Inverse Problem in Potential Scattering, Bulgarska Akad. Nauk. Comptes Rend., 17, 1005-1008, 1964.

Blazek, M., On a Method for Solving the Inverse Problem in Potential Scattering, Commun. Math. Phys., 3, 282-291, 1966.

Bloch, A. Sh., On the Determination of a Differential Equation by its Spectral Matrix Functions, Dokl. Akad. Nauk. SSSR, 92, 209-212, 1953.

Bolsterli, M., and MacKenzie, J., Determination of Separable Potential From Phase Shift, Physics, 2, 141-149, 1965.

Borg, G., Eine Umkehrung der Sturm-Liouvilleschen Eigenwertaufgabe, Acta. Math., 78, 1-96, 1946.

Borg, G., "Uniqueness Theorems in the Spectral Theory of

$$\left\{ -\frac{d^2}{dx^2} + V(x) \right\} u = \lambda u'',$$

Eleventh Congress of Scandinavian Mathematics, Trondheim, 276-287, August 22-25, 1949.

Bosco, B., Dispersion Relations for Scattering Wave Functions in Potential Theory, Nuovo Cimento, 17, 558-579, 1960.

Bosco, B., and Sucher, J., Analyticity in the Coupling Constant and Bound States in Potential Theory, Nuovo Cimento, 19, 1183-1188, 1961.

Brown, R. H., and Twiss, R. Q., The Question of Correlation Between Photons in Coherent Light Rays, Nature, 168, 1447-1450, 1956.

Buck, U., and Pauly, H., Determination of Intermolecular Potentials by Inversion of Molecular Beam Scattering Data, J. Chem. Phys., 51, 1662-1664, 1969.

Burdet, G., Dufour, J., and Giffon, M., Completeness Property of Solutions of Schrödinger's Equation in the Complex Angular Momentum, C. R. Acad. Sci. Paris, 259, 2369-2371, 1964.

Burdet, G., and Giffon, M., Sur le probleme de la Construction du Potential a partir des donnees dans le plan du moment angulaire complexe, C. R. Acad. Sci. Paris, 259, 3190-3192, 1964.

Burdet, G., and Giffon, M., A Proof of the Completeness of the Solutions of the Schrödinger Equation in the λ -Plane, Acta Phys. Hung., 19, 263-267, 1965.

Burdet, G., Giffon, M., and Predazzi, E., On the Inversion Problem in the λ -Plane, I, Nuovo Cimento, **36**, 1337-1347, 1965.

Burdet, G., Giffon, M., and Goldberg, J., On the Inversion Problem in the λ -Plane, II, Nuovo Cimento, **44A**, 138-146, 1966.

Buslaev, V. S., Trace Formulas for Schrödinger's Operator in Three-Space, Dokl. Akad. Nauk SSSR, **143**, 1067-1070, 1962 (Soviet Phys. Dokl., **7**, 295).

Buslaev, V. S., Trace Formulas and Some Asymptotic Estimates of the Resolvent Kernel of the Three-Dimensional Schrödinger Equation, in Topics in Mathematical Physics, Vol. 1, Spectral Theory and Wave Processes, M. Sh. Birman, ed., Leningrad University Press, 1966 (English transl.: Consultants Bureau, Plenum Publishing Corp., New York, 1967).

Buslaev, V. S. and Faddeev, L. D., Formulas for Traces for a Singular Sturm-Liouville Differential Operator, Dokl. Akad. Nauk SSSR **132**, 13-16, 1960 (Soviet Math. Dokl., **1**, 451).

Buslaev, V. S., and Merkur'ev, S. P., Trace Equation for a Three-Particle System, Dokl. Akad. Nauk SSSR, **189**, 269-272, 1969 (Soviet Phys. Dokl., **14**, 1055-1057, 1970).

Cabayan, H. S., and Belford, G. G., On Computing a Stable, Least Squares Solution to the Inverse Problem for a Planar Newtonian Potential, SIAM J. Appl. Math., **20**, 51-61, Jan. 1971.

Calogero, F., Equivalence Between Velocity-Dependent Potentials and Static Potentials, Lett. Nuovo Cimento, **2**, 553-560, 1969.

Calogero, F., Corbella, O. D., Degasperis, A., and deStefano, M. B., Variational Bounds for the Potential in Terms of the S-Wave Phase Shift, J. Math. Phys., **9**, 1002-1006, 1968.

Calogero, F., and Cox, J. R., Evaluation of the Potential and the S-Wave Scattering Amplitude from the Discontinuity of the Latter Across its Left-Hand Cut, Nuovo Cimento, **55A**, 786-808, 1968.

Calogero, F., and Degasperis, A., Values of the Potential and its Derivatives at the Origin in Terms of the S-Wave Phase Shift and Bound-State Parameters, J. Math. Phys., **9**, 90-116, 1968.

Chadan, K., On the Connection Between the S-Matrix and a Class of Non-Local Interactions, Nuovo Cimento, **10**, 892-908, 1958.

Chadan, K., The Left-Hand Cut Discontinuity and Equivalent Potentials, Nuovo Cimento, **24**, 379-384, 1962.

Chadan, K., On a Class of Completely Transparent Nonlocal Two-Body Potentials, Nuovo Cimento, **47A**, 510-525, 1967.

Child, M. S., Repulsive Potential Curves from Predissociation Data, J. Molec. Spectr., **33**, 487-493, 1970.

Chudov, L. A., The Inverse Sturm-Liouville Problem, Nat. Sbornik **25**, 451-456, 1949.

Chudov, L. A., A New Variant of an Inverse Sturm-Liouville Problem on a Finite Interval, Dokl. Akad. Nauk SSSR, **109**, 40-43, 1956.

- Colton, D., On the Inverse Scattering Problem for Axially Symmetric Solutions of the Helmholtz Equation, Quart. J. Math. (Oxford), 22, 125-130, 1971.
- Corbella, O. D., Values of the Potential and Its Derivatives at the Origin for a p-Wave Schrödinger Equation IMAF Rep., Universidad Nacional de Cordoba, Cordoba, Argentina, 1969.
- Corbella, O. D., Inverse Scattering Problem for Dirac Particles. Explicit Expressions for the Values of the Potentials and Their Derivatives at the Origin in Terms of the Scattering and Bound State Data, J. Math. Phys., 11, 1695-1713, 1970.
- Corbella, O. D., L-Wave Inverse Scattering Problem, J. Math. Phys., 12, 1873-1882, Sept. 1971.
- Corinaldesi, E., Construction of Potentials from Phase Shifts and Binding Energies of Relativistic Equations, Nuovo Cimento, 11, 468-478, 1954 and Ibid 12, 469, 1954.
- Coudray, C., and Coz, M., Generalized Translation Operators and the Construction of Potentials at Fixed Energy, Ann. Phys., 61, 488-529, 1970.
- Coudray, C., and Coz, M., Construction of Relativistic Potentials When the Energy is Fixed, J. Math. Phys., 12, 1166-1178, July 1971.
- Cox, J. R., Construction of Potentials from the Many-Channel S-Matrix, Ph.D. Thesis, Indiana Univ., 1962.
- Cox, J. R., Many-Channel Bargmann Potentials, J. Math. Phys., 5, 1065-1069, 1964.
- Cox, J. R., and Thompson, K. W., Some Exact Solution of the Schrödinger Equation at Fixed Energy, Bull. Am. Phys. Soc., 14, 579, 1969.
- Cox, J. R., and Thompson, K. W., On the Inverse Scattering Problem at Fixed Energy for Potentials having Non-Vanishing First Moments, J. Math. Phys., 11, 805-814, 1970a.
- Cox, J. R., and Thompson, K. W., Note on the Uniqueness of the Solution of an Equation of Interest in the Inverse Scattering Problem, J. Math. Phys., 11, 815-817, 1970b.
- Crichton, J. H., Phase-Shift Ambiguities for Spin-Independent Scattering, Nuovo Cimento, 45A, 256-258, 1966.
- Crum, M. M., Associated Sturm-Liouville Systems, Quart. J. Math., 6, 121-127, 1955.
- Degasperis, A., On the Inverse Problem for the Klein-Gordon S-Wave Equation, J. Math. Phys., 11, 551-567, 1970.
- Dikii, L. A., Trace Formulas for Sturm-Liouville Differential Operators, Usp. Matem. Nauk, 13, 111, 1958 (Transl. Amer. Math. Soc., 18, 81, 1961).
- Eftimiu, C., Approximate Solutions of the Unitarity Equation, J. Math. Phys., 11, 2070-2074, 1970.
- Eftimiu, C., An Exact Solution of the Unitarity Equation, Lett. Nuovo Cimento, 4, 475-476, 1970.

- Eftimiu, C., On the Martin-Newton Condition for Finite-Dimensional Unitary Matrices, J. Math. Phys., 12, 2047-2049, Oct. 1971.
- Faddeev, L. D., Uniqueness of Solutions of the Inverse Scattering Problem, Vestnik Leningrad Univ., 11, 126-130, 1956.
- Faddeev, L. D., An Expression for the Trace of the Difference Between Two Singular Differential Operators of the Sturm-Liouville Type, Dokl. Akad. Nauk. SSSR, 115, 878-881, 1957.
- Faddeev, L. D., On the Relations Between S-Matrix and Potential for the One-Dimensional Schrödinger Operator, Dokl. Akad. Nauk. SSSR, 121, 63-66, 1958.
- Faddeev, L. D., The Inverse Problem in the Quantum Theory of Scattering, Usp. Mat. Nauk. (N.S.), 14, 4, 57-119, 1959, (J. Math. Phys., 4, 72-104, 1963).
- Faddeev, L. D., Properties of the S-Matrix of the One-Dimensional Schrödinger Equation, Tr. Mater. Inst. Steklov., 73, 314-336, 1964 (Am. Math. Soc. Transl., 65, 139-166).
- Faddeev, L. D., Increasing Solutions of the Schrödinger Equation, Dokl. Akad. Nauk. SSSR, 165, 514-517, 1965 (Soviet Physics, Doklady, 10, 1033-1035, 1966).
- Faddeev, L. D., Factorization of the S-Matrix for the Multidimensional Schrödinger Equation, Dokl. Akad. Nauk. SSSR, 167, 69-72, 1966 (Soviet Physics, Dokl., 11, 209-211, 1966).
- Fiedeldey, H., The Inverse Scattering Problem for Rank-Two Separable Potentials and Off-Shell Effects, Nucl. Phys., A135, 353-377, 1969.
- Fiedeldey, H., The Dependence of the Triton Binding Energy on the High Energy Phase Shift, Nucl. Phys., A156, 242-256, 1970.
- Firsov, O. B., Determination of Forces Acting Between Atoms with the Use of the Differential Cross-Section of Elastic Scattering, Zh. Eksp. Teor. Fiz., 24, 279-283, 1953.
- Friedman, A., On the Properties of a Singular Sturm-Liouville Equation Determined by its Spectral Function, Mich. Math. J., 4, 137-145, 1957.
- Froberg, C. E., Calculation of the Interaction Between Two Particles From the Asymptotic Phase, Phys. Rev., 72, 519-520, 1947.
- Froberg, C. E., Calculation of the Potential From the Asymptotic Phase, Ark. Mat. Astron. Fys., 34A, 1-16, 1948a.
- Froberg, C. E., Calculation of the Potential From the Asymptotic Phase. Part II, Ark. Mat. Astron. Fys., 36A, 1-55, 1948b.
- Froberg, C. E., On the Determination of Proton-Proton Interaction From Scattering Experiments, Ark. Fys., 3, 1-34, 1951.
- Fuda, M. G., Off-Shell T-Matrix and the Jost Function, Phys. Rev., C1, 1910-1924, 1970.
- Fulton, T., and Newton, R. G., Explicit Non-Central Potentials and Wave Functions for Given S-Matrices, Nuovo Cimento, 10, 677-717, 1956.

Gardner, C. S., Green, J. M., Kruskal, M. D., and Miura, R. M., Methods for Solving the Korteweg-de Vries Equation, Phys. Rev. Lett., 19, 1095-1097, 1967.

Gasymov, M. G., Analytic Properties of the Spectral Function of a Sturm-Liouville Self-Adjoint Operator, Dokl. Akad. Nauk. SSSR, 150, 971-974, 1963 (Soviet Math., 4, 780-783).

Gasymov, M. G., The Inverse Scattering Problem for a System of Dirac Equations of Order $2n$, Dokl. Akad. Nauk. SSSR, 169, 1037-1040, 1966 (Soviet Phys. Dokl., 11, 676).

Gasymov, M. G., An Inverse Problem of Scattering Theory for a System of Dirac Equations of Order $2n$, Tr. Moskov. Mat. Obsc. 19, 41-112, 1968 (Math. Rev., 39, 2418).

Gasymov, M. G., and Levitan, B. M., The Inverse Problem for a Dirac System, Dokl. Akad. Nauk. SSSR, 167, 967-970, 1966 (Soviet Math., 7, 495).

Gel'fand, I. M., and Levitan, B. M., On the Determination of a Differential Equation by its Spectral Function, Dokl. Akad. Nauk. USSR, 77, 557-560, 1951a.

Gel'fand, I. M., and Levitan, B. M., On the Determination of a Differential Equation From its Spectral Function, Isv. Akad. Nauk. SSSR Ser. Mat., 15, 309-360, 1951b (Am. Math. Soc. Transl., 1, 253-304).

Gerber, R. B., and Karplus, M., Determination of the Phase of the Scattering Amplitude From the Differential Cross Section, Phys. Rev., D1, 998-1012, 1970.

Gerber, R. B., and Karplus, M., On the Determination of the Phases of Electromagnetic Scattering Amplitudes From Experimental Data, J. Chem. Phys., to be published, 1971.

Gohberg, I. C., and Krein, M. G., Systems of Integral Equations on the Half-Line with Kernels Depending on the Difference of the Arguments, Usp. Mat. Nauk. (N.S.), 13, 2, 80, 3-72, 1958.

Goldberg, H., Remark on the Martin-Newton Condition for Solubility of the Elastic Unitarity Integral Equation, Phys. Rev., D1, 1242, 1970.

Goldberger, M. L., Lewis, H. W., and Watson, K. M., Use of Intensity Correlations to Determine the Phase of a Scattering Amplitude, Phys. Rev., 132, 2764-2787, 1963.

Goldberger, M. L., Lewis, H. W., and Watson, K. M., Intensity-Correlation Spectroscopy, Phys. Rev., 142, 25-32, 1966.

Goldberger, M. L., and Watson, K. M., Measurement of Time Correlations for Quantum Mechanical Systems, Phys. Rev., 134, B 919-928, 1964.

Goldberger, M. L., and Watson, K. M., Fluctuations with Time of Scattered-Particle Intensities, Phys. Rev., 137, B 1396-1409, 1965a.

Goldberger, M. L., and Watson, K. M., Accuracy of Measurement for Counting and Intensity-Correlation Experiments, Phys. Rev., 140, B 500-509, 1965b.

Gourdin, M., and Martin, A., Interaction non locale separable et matrice de collision, Nuovo Cimento, 6, 757-779, 1957.

- Gourdin, M., and Martin, A., Exact Determination of a Phenomenological Separable Interaction, Nuovo Cimento, 8, 699-707, 1958.
- Gugushvili, E. I., and Mentkovsky, Yu. L., The Riemann Method in the Inverse Problem of Scattering Theory, I & II, Kiev preprints, 1970.
- Holmberg, B., A Remark on the Uniqueness of the Potential Determined From the Asymptotic Phase, Nuovo Cimento, 9, 597-604, 1952.
- Hodgson, P. E., The Optical Model of Elastic Scattering, Oxford at the Clarendon Press (Great Britain), 1963.
- Hooshyar, M. A., On the Inverse Scattering Problem at Fixed Energy for Tensor and Spin-Orbit Potentials, J. Math. Phys., 12, 2243-2258, Oct. 1971.
- Hylleraas, E. A., Calculation of a Perturbing Central Field of Force from the Elastic Scattering Phase Shift, Phys. Rev., 74, 48-51, 1948.
- Hylleraas, E. A., Determination of a Perturbing Potential from its Scattering Phase Shift and Bound State Energy Levels, Nucl. Phys., 57, 208-231, 1964.
- Jauho, P., On the Unique Determination of the Nuclear Potential Between Charged Nucleons with the Aid of Scattering Experiments, Ann. Acad. Sci. Fennicae, Ser. A, 1-43, 1950.
- Jost, R., Über die falschen Nullstellen der Eigenwerte der S-Matrix, Helv. Phys. Acta, 20, 256-266, 1947.
- Jost, R., Eine Bemerkung über den Zusammenhang von Streuphase und Potential, Helv. Phys. Acta, 29, 410-418, 1956.
- Jost, R., and Kohn, W., Construction of a Potential from a Phase Shift, Phys. Rev., 87, 977-922, 1952a.
- Jost, R., and Kohn, W., Equivalent Potentials, Phys. Rev., 88, 382-385, 1952b.
- Jost, R., and Kohn, W., On the Relation Between Phase Shift, Energy Levels and the Potential, Kgl. Dan. Vidensk. Selsk. Mat. Fys. Medd., 27, 1-19, 1953.
- Kac, I. S., Some Cases of Uniqueness of the Solution of the Inverse Problem for Strings with a Boundary Condition at the Singular End, Dokl. Akad. Nauk. SSSR, 164, 975-978, 1965 (Soviet Math., 6, 1324-1327).
- Kay, I., The Inverse Scattering Problem, Research Report No. EM-74, New York Univ., Inst. Math. Sci., Div. Electromagn. Res., 1955.
- Kay, I., The Inverse Scattering Problem When the Reflection Coefficient is a Rational Function, Commun. Pure Appl. Math., 13, 371-393, 1961.
- Kay, I., and Moses, H. E., The Determination of the Scattering Potential from the Spectral Measure Function, I. Continuous Spectrum, Nuovo Cimento, 2, 917-961, 1955.
- Kay, I., and Moses, H. E., The Determination of the Scattering Potential from the Spectral Measure Function, II. Point Eigenvalues and Proper Eigenfunctions, Nuovo Cimento, 3, 66-84, 1956a.

- Kay, I., and Moses, H. E., Reflectionless Transmission Through Dielectrics and Scattering Potentials, J. Appl. Phys., 27, 1503-1508, 1956b.
- Kay, I., and Moses, H. E., The Determination of the Scattering Potential from the Spectral Measure Function, III. Calculation of the Scattering Potential from the Scattering Operator for the One-Dimensional Schrödinger Equation, Nuovo Cimento, 3, 276-304, 1956c.
- Kay, I., and Moses, H. E., The Determination of the Scattering Potential from the Spectral Measure Function, IV. 'Pathological' Scattering Problems in One Dimension, Nuovo Cimento Suppl., 5, 230-242, 1957.
- Kay, I., and Moses, H. E., A Simple Verification of the Gel'fand-Levitan Equation for the Three-Dimensional Scattering Problem, Commun. Pure Appl. Math., 14, 435-445, 1961a.
- Kay, I., and Moses, H. E., The Determination of the Scattering Potential from the Spectral Measure Function. V. The Gel'fand-Levitan Equation for the Three-Dimensional Scattering Problem, Nuovo Cimento, 22, 689-705, 1961b.
- Keller, J. B., Kay, I., and Shmoys, J., Determination of the Potential from Scattering Data, Phys. Rev., 102, 557-559, 1956.
- Keller, J. B., and Zumino, B., Determination of Intermolecular Potentials from Thermodynamic Data and The Law of Corresponding States, J. Chem. Phys., 30, 1351-1353, 1959.
- Keller, J. B., Determination of a Potential from its Energy Levels and Undetectability of Quantization at High Energy, Amer. J. Phys., 30, 22-26, 1962.
- Kostarev, A. A., Perturbation Theory in Inverse Scattering, Izv. VUS Fiz., 4, 109-112, 1964.
- Krein, M. G., The Ideas of P. L. Cebsev and A. A. Markov on the Theory of Limiting Values of Integrals and Their Further Development, Usp. Matem. Nauk, 6, 3-120, 1951a.
- Krein, M. G., Solution of the Inverse Sturm-Liouville Problem, Dokl. Akad. Nauk. SSSR, 76, 21-24, 1951b.
- Krein, M. G., On the Transfer Function of a One-Dimensional Boundary Problem of the Second Order, Dokl. Akad. Nauk. SSSR, 88, 405-408, 1953a.
- Krein, M. G., On Some Cases of Effective Determination of the Density of an Inhomogeneous Cord from its Spectral Function, Dokl. Akad. Nauk. SSSR, 93, 617-620, 1953b (Transl. by Morris B. Friedman, 2 Pine St., West Concord, Mass., 1955).
- Krein, M. G., On a Method of Effective Solution of an Inverse Boundary Problem, Dokl. Akad. Nauk. SSSR, 94, 987-990, 1954a.
- Krein, M. G., On Integral Equations Generating Differential Equations of 2nd Order, Dokl. Akad. Nauk. SSSR, 97, 21-24, 1954b.
- Krein, M. G., On the Determination of the Potential of a Particle From Its S-Function, Dokl. Akad. Nauk. SSSR, 105, 433-436, 1955.
- Krein, M. G., On the Theory of Accelerants and S-Matrices of Canonical Differential Systems, Dokl. Akad. Nauk. SSSR, 111, 1167-1170, 1956.

Krein, M. G., On the Continual Analogue of a Christoffel Formula for the Theory of Orthogonal Polynomials, Dokl. Akad. Nauk. SSSR, 113, 970-973, 1957.

Krein, M. G., Integral Equations on the Half-Line with a Kernel Depending on the Difference of the Arguments, Usp. Mat. Nauk., 13, 5, 83, 3-120, 1958.

Landau, L. D., and Lifshits, E. M., Statistical Physics, Pergamon Press, London, 1959.

Lavrent'ev, M. M., On an Inverse Problem for the Wave Equation, Dokl. Akad. Nauk. SSSR, 157, 520-521, 1964 (Soviet Math., 5, 970-972).

Lavrent'ev, M. M., A Class of Inverse Problems for Differential Equations, Dokl. Akad. Nauk. SSSR, 160, 32-35, 1965 (Soviet Math., 6, 29-32).

Lax, P. D., and Phillips, R. S., Scattering Theory, Bull. Amer. Math. Soc., 70, 130-142, 1964.

Lax, P. D., and Phillips, R. S., Analytic Properties of the Schrödinger Scattering Matrix, in Perturbation Theory, C. H. Wilcox, ed., Wiley, 1966.

Levinson, N., On the Uniqueness of the Potential in a Schrödinger Equation for a Given Asymptotic Phase, Kgl. Dan. Vidensk. Selsk. Mat-Fys. Medd., 25, 1-29, 1949. a.

Levinson, N., Determination of the Potential from the Asymptotic Phase, Phys. Rev., 75, 1445, 1949. b.

Levinson, N., Certain Explicit Relationships Between Phase Shifts and Scattering Potential, Phys. Rev., 89, 755-757, 1953.

Levitan, B. M., Certain Questions in the Spectral Theory of Self-Adjust Differential Operators, Usp. Mat. Nauk. (N.S.), 11, 6, 72, 117-144, 1956 (Amer. Math. Soc. Transl., 18, 49).

Levitan, B. M., and Sargsyan, I. S., Some Problems in the Theory of Sturm-Liouville Equations, Usp. Mat. Nauk. (N.S.), 15, 1, 3-98, 1960 (Russian Math. Surveys, 15, 1-95).

Levitan, B. M., On the Determination of a Differential Equation by Two of its Spectra, Dokl. Akad. Nauk. SSSR, 150, 474-476, 1963 (Soviet Math., 4, 691-693).

Levitan, B. M., On the Determination of a Differential Equation by Two of its Spectra, Izv. Akad. Nauk. SSSR, 28, 63-78, 1964. a.

Levitan, B. M., Generalized Translation Operators and Some of Their Applications, Translated from the Russian by the Israel Program for Scientific Translations, Jerusalem, 1964. b. (published in the USA by Daniel Davey and Co., Inc., N.Y.).

Levitan, B. M., and Gasymov, M. G., Determination of a Differential Equation by Two of its Spectra, Russian Math. Surveys, 19, 1-63, 1964.

Ljance, V. E., A Differential Operator with Spectral Singularities, I, II, Mat. Sb., 64, 521-561; 65, 47-103, 1964 (Amer. Math. Soc. Transl. ser 2, 60, 185-283).

Ljance, V. E., The Inverse Problem for a Nonselfadjoint Operator, Dokl. Akad. Nauk. SSSR, 166, 30-33, 1966 (Soviet Math., 7, 27-30).

- Loeffel, J. J., On an Inverse Problem in Potential Scattering Theory, Ann. Inst. H. Poincaré, 8, 339-347, 1968.
- Mal'cenko, V. I., The Inverse Problem for the Equations of Quantum Mechanics with Energy-Dependent Potentials, Ukr. Mat. Zh., 18, 126-129, 1966.
- Marchenko, V. A., Some Problems of the Theory of Second-Order Differential Operators, Dokl. Akad. Nauk. SSSR, 72, 457-460, 1950.
- Marchenko, V. A., Some Problems in the Theory of One-Dimensional Second-Order Differential Operators, part I, Tr. Mosk. Mat. Obs., 1, 327-420, 1952; and *ibid.*, 2, 3-82, 1953.
- Marchenko, V. A., On the Reconstruction of the Potential Energy from Phases of the Scattered Waves, Dokl. Akad. Nauk. SSSR, 104, 695-698, 1955.
- Marchenko, V. A., Expansion in Eigenfunctions of Non-Self-Adjoint Singular Differential Operators of Second Order, Mat. Sbornik (N.S.), 52, 739-788, 1960 (Amer. Math. Soc. Transl., 25, 77).
- Martin, A., For the Generalization of Levinson Theorem to Nonlocal Interactions, Nuovo Cimento, 7, 607, 1958.
- Martin, A., S-Matrix, Left-Hand Cut Discontinuity and Potential, Nuovo Cimento, 19, 1257-1265, 1961.
- Martin, A., Construction of the Scattering Amplitude From the Differential Cross Sections, Nuovo Cimento, 59A, 131-152, 1969.
- Martin, A., and Targonski, A., On the Uniqueness of a Potential Fitting a Scattering Amplitude at a Given Energy, Nuovo Cimento, 20, 1182-1190, 1961.
- Mason, E. A., and Monchick, L., Methods for the Determination of Intermolecular Forces, Advan. Chem. Phys., 12, 350-387, 1967.
- McVoy, K. W., Heller, L., and Bolsterli, M., Optical Analysis of Potential Well Resonances, to be published; 1971.
- Melkanoff, M. A., Saxon, D. S., and Cantor, D. G., A Fortran Program for Elastic Scattering Analyses with the Nuclear Optical Model, Univ. California Press, 1961.
- Melnikov, V. K., On Approximate Methods in the Inverse Problem of the Quantum Theory of Scattering, Usp. Mat. Nauk. (N.S.), 14, 4 (88), 121-131, 1959, (Amer. Math. Soc. Transl., 25, 271-282).
- Miller, K., and Viano, G. A., On the Stability of the Complex Angular Momentum Continuation, Nucl. Phys., B25, 460-470, 1971.
- Miller, W. H., WKB Solution of Inversion Problems for Potential Scattering, J. Chem. Phys., 51, 3631-3638, 1969.
- Miller, W. H., Additional WKB Inversion Relations for Bound-State and Scattering Problems, J. Chem. Phys., to be published, 1971.

- Mills, R. L., and Reading, J. F., Inversion Problem with Separable Potentials, J. Math. Phys., 10, 321-331, 1969.
- Moses, H. E., Calculation of the Scattering Potential From Reflection Coefficients, Phys. Rev., 102, 559-567, 1956.
- Mushkhelishvili, N. I., Singular Integral Equations, Groningen, Holland, 1953.
- Neigauz, M. G., About the Determination of the Asymptotic Behavior of the Function $q(x)$ on the Basis of the Spectral Function of $-y'' + q(x)y$, Dokl. Akad. Nauk. SSSR, 102, 25-28, 1955.
- Newton, R. G., Connection Between the S-Matrix and the Tensor Force, Phys. Rev., 100, 412-428, 1955.
- Newton, R. G., Remarks on Scattering Theory, Phys. Rev., 101, 1588-1596, 1956.
- Newton, R. G., Electron Scattering by the Deuteron, Phys. Rev., 105, 763-764, 1957.
- Newton, R. G., Analytic Properties of Radial Wave Functions, J. Math. Phys., 1, 319-347, 1960.
- Newton, R. G., Construction of Potentials from the Phase Shifts at Fixed Energy, J. Math. Phys., 3, 75-82, 1962.
- Newton, R. G., Scattering Theory of Waves and Particles, McGraw-Hill, New York, 1966.
- Newton, R. G., Connection Between Complex Angular Momenta and the Inverse Scattering Problem at Fixed Energy, J. Math. Phys., 8, 1566-1570, 1967.
- Newton, R. G., Determination of the Amplitude from the Differential Cross Section by Unitarity, J. Math. Phys., 9, 2050-2055, 1968.
- Newton, R. G., Inverse Problems in Physics, SIAM Rev., 12, 346-356, 1970.
- Newton, R. G., and Fulton, T., Phenomenological Neutron-Proton Potentials, Phys. Rev., 107, 1103-1111, 1957.
- Newton, R. G., and Jost, R., The Construction of Potentials from the S-Matrix for Systems of Differential Equations, Nuovo Cimento, 1, 590-622, 1955.
- O'Brien, T. J. P., and Bernstein, R. B., Investigation of the Hylleraas Method for Determining the Potential-Energy Function From the Phase Shift, J. Chem. Phys., 51, 5112-5117, 1969.
- Ohmura, T., Extensions of Variational Methods, III-Determination of Potential From Phase Shift Function, Progr. Theor. Phys., 16, 231-243, 1956.
- Pearce, W. A., A Note on the Ohmura Method of Potential Deduction, Progr. Theor. Phys., 32, 180-181, 1964.
- Percival, I. C., Energy Moments of Scattering Phase Shifts, Proc. Phys. Soc. (London), 80, 1291-1300, 1962.
- Percival, I. C., and Roberts, A. J., Energy Moments of Scattering Phase Shifts, II. Higher Partial Waves, Proc. Phys. Soc. (London), 82, 519-528, 1963.

- Petrás, M., Singularities of Jost Functions and Potentials, Czech. J. Phys., B12, 87-92, 1962.
- Portinari, J. C., The One-Dimensional Inverse Scattering Problem, Ph.D. Thesis, MIT Department of Electrical Engineering, June 1966.
- Portinari, J. C., Finite Range Solutions to the One-Dimensional Inverse Scattering Problem, Ann. Phys., 45, 445, 1967.
- Prats, F., and Toll, J. S., Construction of the Dirac Equation Central Potential from Phase Shifts and Bound States, Phys. Rev., 113, 363-370, 1959.
- Prosser, R. T., Formal Solutions of Inverse Scattering Problems, J. Math. Phys., 10, 1819-1822, 1969.
- Ramm, A. G., Conditions Under Which Scattering Matrices are Analytic, Dokl. Akad. Nauk. SSSR, 157, 1073-1076, 1964 (Soviet Phys. Dokl., 9, 645-647).
- Redmond, P. J., Some Remarks Concerning a Pathological Matrix of Interest in the Inverse-Scattering Problem, J. Math. Phys., 5, 1547-1554, 1964.
- Regge, T., Construction of Potentials from Resonance Parameters, Nuovo Cimento, 9, 491-503, 1958.
- Regge, T., Introduction to Complex Orbital Momenta, Nuovo Cimento, 14, 951-976, 1959.
- Regge, T., Bound States, Shadow States and Mandelstam Representation, Nuovo Cimento, 18, 947-956, 1960.
- Roberts, M. J., Energy Moments and Cross Sections for a Gaussian Potential, Proc. Phys. Soc. (London), 82, 594-604, 1963.
- Roberts, M. J., Negative Energy Moments and Low Energy Approximations to Phase Shifts, Proc. Phys. Soc. (London), 83, 503-517, 1964.a.
- Roberts, M. J., Alternative Derivation of the Moment Relations Involving Scattering Phase Shifts, Proc. Phys. Soc. (London), 84, 825-826, 1964.b.
- Roberts, M. J., Moment Relations for Schrödinger Two-Body Cross Sections, Proc. Phys. Soc. (London), 86, 683-691, 1965.
- Roberts, R. E., and Ross, J., Potential Inversion for the Semiclassical Optical Model, J. Chem. Phys., 53, 2126-2127, 1970.
- Roman, P., and Marathay, A. S., Analyticity and Phase-Retrieval, Nuovo Cimento, 30, 1452-1464, 1963.
- Ross, M., ed., Quantum Scattering Theory, Indiana University Press, Bloomington, Ind., 1963.
- Sabatier, P. C., Interpolation des Fonction d'Onde dans le Plan Complexe du Moment Angulaire, C.R. Acad. Sci., Paris, Ser. A., 263, 788-790, 1966.a.
- Sabatier, P. C., Asymptotic Properties of the Potentials in the Inverse Scattering Problem at Fixed Energy, J. Math. Phys., 7, 1515-1531, 1966.b.

Sabatier, P. C., Analytic Properties of a Class of Potentials and the Corresponding Jost Functions, J. Math. Phys., 7, 2079-2091, 1966.c.

Sabatier, P. C., Le problème inverse à énergie fixée en mécanique quantique, These de Doctorat d'Etat: Université Paris-Orsay, Mars 1966.d.

Sabatier, P. C., General Method for the Inverse Scattering Problem at Fixed Energy, J. Math. Phys., 8, 905-918, 1967.a.

Sabatier, P. C., Interpolation Formulas in the Angular Momentum Plane, J. Math. Phys., 8, 1957-1972, 1967.b.

Sabatier, P. C., Approximation des Potentiels par les Potentiels de Classe, C.R. Acad. Sc., Paris, 265, 5-8, 1967.c.

Sabatier, P. C., Approach to Scattering Problems Through Interpolation Formulas and Application to Spin-Orbit Potentials, J. Math. Phys., 9, 1241-1258, 1968.

Sabatier, P. C., A New Tool for Scattering Studies, J. Math. Phys., 12, 1303-1326, July 1971.

Sabatier, P. C., Complete Solution of the Inverse Scattering Problem at Fixed Energy, J. Math. Phys., to be published, 1972.

Sabatier, P. C., and Quyen van Phu, F., Numerical Computations in the Inverse Scattering Problem at Fixed Energy, Phys. Rev., D4, 127-132, 1971.

Schonbeck, T. P., On Inverse Scattering for the Klein-Gordon Equation, preprint, Florida Atlantic Univ., Dept. of Math.

Sprung, D. W. L., and Srivastava, M. K., Soft-Core Potential Model for Nucleon-Nucleon Scattering. II, Nucl. Phys., A139, 605-624, 1969.

Srivastava, M. K., Off-Shell Behavior of Phase-Shift Equivalent Momentum Dependent Potentials, Nuclear Phys., A157, 61-72, 1970.

Srivastava, M. K., Banerjee, P. K., and Sprung, D. W. L., A Very Soft Core Potential Model for 1S_0 Neutron-Proton Scattering, Phys. Lett., 29B, 635-637, 1969.

Stashevskaya, V. V., On Inverse Problems of Spectral Analysis for a Class of Differential Equations, Dokl. Akad. Nauk. SSSR, 93, 409-411, 1953.

Swan, P., and Pearce, W. A., Deduction of Potentials from Scattering Phase Shifts-I) Neutral Particles, Nucl. Phys., 79, 77-107, 1966.

Swan, P., Deduction of Potentials from Scattering Phase Shifts, II. Charged Particle, Nucl. Phys., 90, 436-448, 1967.

Tabakin, F., Inverse Scattering Problem for Separable Potentials, Phys. Rev., 177, 1443-1451, 1969.

Tietz, T., Discrete States for Non-Singular and Singular Potential Problems, Nuovo Cimento, 11, 126-130, 1959.

Titchmarsh, E. C., Theory of Fourier Integrals, Oxford, 1959.

Twiss, R. Q., Little, A. G., and Brown, R. H., Correlation Between Photons, in Coherent Beams of Light, Detected by a Coincidence Counting Technique, Nature, 180, 324-326, 1957.

Uhlenbeck, G. E., and Beth, E., Quantum Theory of the Non-ideal Gas. Part I. Deviations From Classical Theory, Physica, 3, 729-749, 1936.

Verde, M., Asymptotic Expansions of Phase Shifts at High Energies, Nuovo Cimento, 2, 1001-1014, 1955.

Verde, M., The Inversion Problem in Wave Mechanics and Dispersion Relations, Nuclear Phys., 9, 255-266, 1958/59.

Viano, G. A., A Note on Inverse Problems in Potential Scattering, Nuovo Cimento, 63, 581-597, 1969.

Volk, V. Ya., On Inverse Formulas for a Differential Equation with a Singularity at $x = 0$, Usp. Mat. Nauk. (N.S.), 8, 4 (56), 141-151, 1953.

Vollmer, G., Inverse Problem in Atom Scattering in WKB Approach, Z. Phys., 226, 423-434, 1969.

Volterra, V., Theory of Functionals and Integro-Differential Equations, Dover Publications, New York, 1959.

Weiss, R., and Scharf, G., The Inverse Problem in Potential Scattering According to the Klein-Gordon Equation, Helv. Phys. Acta, to be published, 1971.

Weiss, R., Stahel, W. and Scharf, G., The Inverse Problem in Potential Scattering According to the Dirac Equation, Univ. Zurich preprint.

Wheeler, J. A., Scattering and Potential, Phys. Rev., 99, 630, 1955.

6. ELECTROMAGNETIC SCATTERING

This chapter is concerned with the broad range of practical inversion problems in the field of electromagnetic scattering; these problems are also discussed by *Mitra, Bojarski, Sondhi, and Prosser* in Chapter 8. Certain of the problems are amenable to full-wave treatment, while others require geometrical or physical optics assumptions. The authors invariably are concerned with several facets of the problem:

1. Characterization of the scattering data. One wishes to know the necessary and sufficient conditions to ensure that scattering amplitude functions really are scattering amplitudes of properly posed problems; such conditions usually are of the form of analyticity requirements for exact problems.
2. Practicality of solutions. Algorithms that invert the scattering data to yield the scatterer function should be simple and "practical"; they should use a minimum of data.
3. Relationship between approximate and exact techniques.
4. Stability criteria. Small errors in scattering amplitudes should not lead to large errors in characterization of the scatterer.
5. Realizability criteria. These are requirements on scattering properties that lead to physical, realizable scatterers.
6. Extension of inversion techniques to dissipative systems.

H. E. Moses organized and chaired the session devoted to electromagnetic scattering.

THE INVERSE SCATTERING PROBLEM FOR TRANSMISSION LINES

Irvin Kay

Wayne State University

ABSTRACT

1 N73-11612

A number of exact and approximate methods for solving the inverse scattering problem for transmission lines are reviewed. In particular, the application to transmission lines of *Marcenko's* [1955] version of the Gelfand-Levitan exact method for the quantum mechanical problem is compared with a more direct approach based on a different version of the Gelfand-Levitan method. In addition, some aspects of the lack of uniqueness of solutions are discussed, and some open questions related to the inverse scattering problem are suggested.

INTRODUCTION

Transmission line scattering is distinguished here by its one-dimensional character and the fact that waves propagate along the line in two directions, from left to right and from right to left. It takes place, typically, in a cylindrically symmetric medium that extends to infinity in the direction of propagation. The physical properties of such a medium are stratified — they vary only in the direction of propagation — although it may be enclosed in a cylindrical structure.

A complete description of electromagnetic wave propagation in a cylindrical structure requires representing the fields as infinite sums of modes. Each field in a mode has a vector factor which depends only on coordinates transverse to the cylindrical axis and a scalar factor which depends only on position along the cylindrical axis. In the lossless case, in each mode a pair of such scalar factors, corresponding to the electric and magnetic fields, satisfies the transmission line equations

$$\frac{dV}{dz} = i\omega L(z)I \quad (1)$$

$$\frac{dI}{dz} = i\omega C(z)V$$

where V is associated with the electric field, I is associated with the magnetic field, ω is the frequency, $L(z)$ is the effective distributed inductance, and $C(z)$ is the effective distributed capacitance [Marcuvitz, 1964, pp. 1-5].

Other than cylindrical structures can lead to a transmission line description of wave propagation, e.g., a spherically symmetric medium [Marcuvitz, 1964, pp. 29-54]. The requirement is that solutions of the basic governing equations be separable into factors, one of which is a function of a single variable whose values denote positions along a propagation path. This variable can represent curvilinear displacement and might, for example, be one of the angular coordinates in a spherical system.

The transmission line equations (1) can be generalized so that the V and I are n component vectors and L and C are $n \times n$ matrices; such a generalization is necessary, for example, when the medium is anisotropic. Then the physical prototype is a system of coupled $2n$ wire transmission lines.

For certain nonseparable cases the transmission line formalism can provide an approximate model for wave propagation. For example, it can be used for a tapered waveguide with slowly varying cross section [Shelkunoff, 1943, pp. 405-406].

The physical meaning of the parameters in Eq. (1) may be modified on occasion. For example, *Becher* and *Sharpe* [1969] have applied the transmission line formalism to the problem of determining the conductivity of the earth from its low frequency propagation effects. In this case, the functions $C(z)$ and $L(z)$ are proportional to the square root of the conductivity and its reciprocal.

DIRECT VS INDIRECT SCATTERING

The direct scattering problem for the transmission line is equivalent to the problem of calculating its input impedance at its terminals [Kay, 1963], which can be taken to be $\pm \infty$ in general. It is assumed that

$$\begin{aligned} \lim_{z \rightarrow -\infty} L(z) &= L_1 & \lim_{z \rightarrow -\infty} C(z) &= C_1 \\ \lim_{z \rightarrow +\infty} L(z) &= L_0 & \lim_{z \rightarrow +\infty} C(z) &= C_0 \end{aligned} \quad (2)$$

where L_1 , C_1 , L_0 , C_0 are all constants.

Under appropriate conditions Eq. (1) also has the property that a solution exists for which the asymptotic behavior of V is

$$\begin{aligned} V &\sim \exp(ik_0 z) + r(\omega)\exp(-ik_0 z) & \text{for } z \rightarrow -\infty \\ V &\sim t(\omega)\exp(ik_1 z) & \text{for } z \rightarrow +\infty \end{aligned} \quad (3)$$

where

$$k_0 = \omega(L_0 C_0)^{1/2} \quad k_1 = \omega(L_1 C_1)^{1/2}$$

Then

$$\begin{aligned} I &\sim Y_0[\exp(ik_0 z) - r(\omega)\exp(-ik_0 z)] & \text{for } z \rightarrow -\infty \\ I &\sim Y_1(\omega)\exp(ik_1 z) & \text{for } z \rightarrow +\infty \end{aligned} \quad (4)$$

where

$$Y_0 = Z_0^{-1} = \left(\frac{C_0}{L_0} \right)^{1/2}$$

$$Y_1 = Z_1^{-1} = \left(\frac{C_1}{L_1} \right)^{1/2}$$

The quantities k_i are propagation constants, Y_i characteristic admittances, Z_i characteristic impedances; $r(\omega)$ is the left amplitude reflection coefficient, and $t(\omega)$ is the right amplitude transmission coefficient.

The asymptotic relations (3) and (4) correspond to a wave of unit amplitude incident from the left, partially reflected, and partially transmitted. If a wave of unit amplitude is incident from the right, a reflected wave with amplitude equal to a reflection coefficient $\rho(\omega)$ and transmitted wave with amplitude equal to a transmission coefficient $\tau(\omega)$ will be generated. The coefficients $r(\omega)$, $\rho(\omega)$, $t(\omega)$, $\tau(\omega)$ are elements of a 2×2 matrix called the scattering matrix. The direct scattering problem can also be formulated as the calculation of the scattering matrix of the transmission line. With this formulation, only the function V need be considered in the solution of Eq. (1).

In the inverse scattering problem some part of the scattering matrix, usually the reflection coefficient $r(\omega)$, is given and it is required to calculate $L(z)$ and $C(z)$ or their equivalents. Most often $r(\omega)$ is a prescribed function of frequency, although in some cases the reflection coefficient is given at a fixed frequency and as a function of some other parameter such as angle of incidence.

The inverse scattering problem for quantum mechanics is similar in that some part of the scattering matrix associated with an unknown potential is given as a function of the energy, and it is required to determine the scattering potential. However, in quantum mechanics the problem is essentially three dimensional and therefore leads most naturally to a somewhat different mathematical formulation, even after separation of variables, than does the transmission line propagation problem. This is also true of certain inverse scattering problems for the electromagnetic field — when a field is scattered in a homogeneous medium by an obstacle of unknown shape and it is required to determine the shape or other geometrical and physical properties of the obstacle from a knowledge of the scattered field at large distances.

It is true that for spherically symmetric potentials the quantum mechanics inverse scattering problem reduces to one for the Schrödinger equation in a single variable, the radial coordinate. However, the problem is still essentially three dimensional and differs from that of the transmission line in that it is usually formulated in terms of the phase shift of a standing wave rather than reflection coefficient of a propagating wave.

For quantum mechanics the ground state radial Schrödinger equation has the form

$$\frac{d^2 \psi}{dr^2} + [E - V(r)] \psi = 0 \quad (5)$$

for the interval

$$0 \leq r < \infty$$

with the boundary condition

$$\psi(0) = 0 \quad (6)$$

and the asymptotic condition

$$\psi \sim A(E) \sin[r + \phi(E)] \text{ as } r \rightarrow \infty \quad (7)$$

The inverse scattering problem is to determine $P(r)$ from a knowledge of $\phi(E)$. The transmission line problem can be transformed into one for a differential equation similar to Eq. (5); however, the interval over which the independent variable ranges is $(-\infty, \infty)$ rather than $(0, \infty)$ and the boundary condition (6) is not imposed.

Mathematically, the difference between the two problems is primarily that the spectrum associated with the quantum mechanics problem is simple, whereas that associated with the transmission line problem has a multiplicity of 2. Because of this difference the Gelfand-Levitan method of solving the inverse scattering problem of quantum mechanics [Gelfand, and Levitan, 1951] must be modified when it is applied to the transmission line problem. The Marcenko variation of the Gelfand-Levitan method [Marcenko, 1955; Agranovich and Marcenko, 1963] was originally applied to the problem of quantum mechanics but has been used by Sharpe [1963] in solving the synthesis problem for transmission lines. This was possible because in the Marcenko method the quantum mechanics model given by Eqs. (5), (6) and (7) is viewed not as a standing wave picture, but as one for a propagating wave incident from the right on a potential with a short circuit terminal impedance condition at 0 and a wave reflected to the right. The reflection coefficient in this case is $\exp(2i\phi)$. When Marcenko's method is applied to the transmission line, however, an unnecessary complication arises because of the artificial boundary condition (Eq. (6)). This will be discussed in more detail later.

The difference in form between Eqs. (1) and (5) is not as crucial as it might seem at first. A transformation exists that will change Eq. (1) into the form of Eq. (5) and thus reduce the inverse scattering problem for the transmission line to one for the Schrödinger equation [Kay and Moses, 1956]. The transformation is given by

$$x = \int_0^z \frac{dz}{v(z)} \quad \Phi = Z^{-1/2} \quad \psi = VZ^{-1/2} \quad (8)$$

where

$$v = (LC)^{-1/2}$$

is the wave velocity and $Z = L/C$ is the characteristic impedance of the medium; then ψ will satisfy Eq. (5) with $P(x)$ related to Φ by

$$\frac{d^2 \Phi}{dx^2} - P(x)\Phi = 0 \quad (9)$$

If the inverse scattering problem for Eq. (5) is solved to determine $P(x)$, Φ can then be determined. However, the range of x in this case is $(-\infty, \infty)$ rather than $[0, \infty)$ as in the problem of quantum mechanics; compare Moses and deRidder, [1963].

The relation between Eqs. (1) and (5) can be more direct as when a plane wave at a fixed frequency is incident on a linearly stratified medium at various angles and the reflection coefficient is given as a function of the angle of incidence rather than frequency. If L is constant and only C varies, L can be eliminated from Eq. (1), and the equation for V becomes

$$\frac{d^2 V}{dz^2} + \left[k_0^2 \cos^2 \alpha - \mu^2(z) \right] V = 0 \quad (10)$$

where k_0 is the wave number in the homogeneous region assumed to be to the left of the inhomogeneous medium, α is the angle of incidence, and $\mu^2(z)$ is the difference between the square of the wave number in the inhomogeneous region and that in the homogeneous region.

QUESTIONS OF UNIQUENESS

In view of Eqs. (8) and (9) the inverse scattering problem for the transmission line clearly does not have a unique solution, since the velocity of propagation $v(z)$ can be any continuous positive function. However, in most propagation media of practical importance the permeability is that of vacuum, and therefore only the permittivity varies. In such a case, there will be a well-defined relation between L and C depending on the particular mode being considered, and this relation will resolve the ambiguity.

Under suitable restrictions a knowledge of the reflection coefficient $r(\omega)$ is sufficient to provide a unique solution to the inverse scattering problem. However, a knowledge of the transmission coefficient $t(\omega)$ is not sufficient to determine a unique solution, in general; compare *Karp and Shmoys* [1955]; *Kay*, [1960]. This fact is implied by a consideration of whether a knowledge of r alone or t alone is sufficient to determine the remaining elements of the scattering matrix. For physically reasonable transmission lines the scattering matrix will be unitary for real ω and the principle of reciprocity will hold. Its elements thus will have the following properties [cf. *Heim and Sharpe*, 1967]:

$$(\text{for } \omega \text{ real}) |t|^2 + |r|^2 = |r|^2 + |\rho|^2 = 1 \quad (\text{conservation of energy})$$

$$t = \tau \quad (\text{reciprocity}) \quad (11)$$

$$(\text{for } \omega \text{ real}) t(-\omega) = t^*(\omega), \quad r(-\omega) = r^*(\omega), \quad \rho(-\omega) = \rho^*(\omega)$$

r, ρ, t are analytic functions of ω , regular in the upper half plane

$$t(\omega) = 1 + o(\omega) \quad \text{for large } \omega \text{ on the real axis and in the upper half plane}$$

$t(\omega)$ has no zeroes in the upper half plane or on the real axis (passive line)

The properties (11) imply that $t(\omega)$ and $\rho(\omega)$ can be determined uniquely from knowledge of $r(\omega)$, but $r(\omega)$, cannot be determined uniquely from a knowledge of $t(\omega)$. In fact, for ω in the upper half plane $t(\omega)$ is given by

$$\log t(\omega) = \frac{1}{2\pi i} \int_{-\infty}^{\infty} \frac{\log[1 - r(\sigma)r(-\sigma)]}{\sigma - \omega} d\sigma \quad (12)$$

When $r(\omega)$ is a rational function of ω

$$t(\omega) = \frac{\prod_{i=1}^n (\omega + \kappa_i)}{\prod_{i=1}^n (\omega - \lambda_i)} \quad (13)$$

where the λ_i are poles of $r(\omega)$ in the lower half plane and the κ_i are the n roots of

$$1 - r(\omega)r(-\omega) = 0 \quad (14)$$

in the upper half plane [Kay, 1966].

These results are based on the assumption that the differential equation does not have a discrete spectrum, which is the case when the scattering matrix is regarded as a function of frequency. However, a discrete spectrum can occur along with the continuous one in the case of a plane wave incident on a linearly stratified medium at a fixed frequency when r is regarded as a function of the angle of incidence so that the differential equation for the problem is Eq. (10). In this case the inverse scattering problem has an infinite number of solutions for a given $r(\omega)$. These occur when the condition that $t(\omega)$ be regular in the upper half plane is violated by the assignment of poles for $t(\omega)$ on the positive imaginary axis. When these poles are not also poles of $r(\omega)$, additional solutions occur. For example, an infinite set of media exists for which the reflection coefficient is identically zero for all angles of incidence [Kay and Moses, 1956b].

EXACT METHODS OF SOLUTION

Exact solutions of the inverse scattering problem have been based on a general technique given by Gelfand and Levitan [1951] for the differential equation (5). This is a method for determining the function $P(r)$ from a knowledge of the spectral measure function $m(E)$ [Titchmarsh, 1946; Newton, 1966, p. 616; Kay and Moses, 1955] in terms of which the spectral resolution associated with the differential operator and the boundary conditions of the scattering problem is defined. In applying the Gelfand-Levitan technique it is usually necessary, as a first step, to construct the spectral measure function from a knowledge of elements of the scattering matrix.

In the case of the quantum mechanics problem for which the given scattering data is the asymptotic phase shift, this is accomplished by the use of a dispersion relation derived from analytic properties of the scattering matrix in the complex plane. For the transmission line problem, where the scattering data is the amplitude reflection coefficient $r(\omega)$, such a procedure is not necessary since in this case the spectral measure function depends only on $r(\omega)$ itself [Kay and Moses, 1956a].

Marcenko [1955] used a modification of the technique to solve the quantum mechanics inverse scattering problem directly in terms of the phase shift. His treatment of the problem can be interpreted as taking the transmission line point of view, wherein a wave incident from $+\infty$ on the right is reflected from a transmission line which is short circuited at 0. The reflected wave, propagating from left to right, has the behavior

$$\psi_{\text{ref}} \sim S(E)\exp(i\sqrt{E}r) \text{ as } r \rightarrow +\infty$$

where for real positive E the factor $S(E)$ has unit amplitude and a phase which is twice that of the phase shift associated with the usual version of the scattering problem. The transmission line reflection coefficient is $S(E)$; the fact that it has unit amplitude is a result of the short circuit at zero and conservation of energy [Kay, 1963].

A solution to the inverse scattering problem associated with Eq. (1) for a short-circuited line has been given by Brown [1966] using Marcenko's method. Sharpe [1963] used Marcenko's method to solve the closely related synthesis problem for Eq. (1) without restricting the problem to a short-circuited line. A direct version of the solution was given by Heim and Sharpe [1967], who were able to avoid the assumption of a short-circuited line at the cost of having to require a knowledge of both the reflection coefficient $r(\omega)$ and the transmission coefficient $t(\omega)$.

A physically oriented, heuristic derivation of a solution to the inverse scattering problem can be given [Kay, 1960] that shows why the transmission line approach leads to a solution depending only on a knowledge of the reflection coefficient. The derivation depends on associating the partial differential equation

$$U_{xx} - P(x)U - U_{tt} = 0 \quad (15)$$

with Eq. (5) and

$$U_{0xx} - U_{0tt} = 0 \quad (16)$$

with the unperturbed form of Eq. (5) — the form which occurs when P vanishes. It is then assumed that U is related to U_0 by a linear transformation which can be expressed in the form of an integral operation on U_0 . The assumption is, explicitly,

$$U(x, t) = U_0(x, t) + \int_{-\infty}^{\infty} K(x, y) U_0(y, t) dy \quad (17)$$

where $K(x, y)$ is the kernel of the integral operator.

If the function $R(\tau)$ is defined by

$$R(\tau) = \frac{1}{2\pi} \int_{-\infty}^{\infty} r(\omega) e^{-i\omega\tau} d\omega \quad (18)$$

$R(x + t)$ will represent the asymptotic reflected wave due to an incident wave $\delta(x - t)$, where δ is the Dirac delta function. The function

$$U_0(x, t) = \delta(x - t) + R(x + t) \quad (19)$$

will then represent the asymptotic form of U when $x \rightarrow -\infty$. The function $R(\tau)$ vanishes for $\tau < 0$ if $r(\omega)$ is analytic and $o(1)$ in the upper half plane. This will be the case when $P(x)$ is zero for negative x .

If $U(x,t)$ that satisfies Eq. (17) is the proper solution to Eq. (15) for the scattering problem it must vanish for $x > t$ since it represents a simple disturbance traveling from left to right with velocity 1. Thus, it can depend on U_0 only at values of x that the disturbance has had time to reach. In view of this appeal to causality the integral in Eq. (17) should only extend from $-\infty$ to x . Then

$$U(x,t) = U_0(x,t) + \int_{-\infty}^x K(x,y)U_0(y,t)dy \quad (20)$$

or, from Eq. (19) and a fundamental property of the delta function,

$$U(x,t) = \delta(x-t) + R(x+t) + K(x,t) + \int_{-\infty}^x K(x,y)R(y+t)dy \quad (21)$$

Now, if $t > x$, Eq. (21) becomes

$$0 = R(x+t) + K(x,t) + \int_{-\infty}^x K(x,y)R(y+t)dy \quad t > x \quad (22)$$

Since R vanishes for negative values of its argument the integral in Eq. (22) actually can be written with $-t$ as its lower limit. Equation (22) is a Fredholm integral equation for the kernel $K(x,t)$ in terms of the Fourier transform of the reflection coefficient $r(\omega)$.

If Eq. (17) is substituted into Eq. (15) it is found, as usual when the Gelfand-Levitan technique is used, that $K(x,t)$ satisfies the differential equation

$$K_{xx} - K_{tt} - P(x)K = 0 \quad (23)$$

subject to the boundary conditions

$$K(x, -\infty) = 0 \quad \frac{2dK(x,x)}{dx} = P(x) \quad (24)$$

the second of which provides the solution of the inverse scattering problem for Eq. (5).

If a Fourier transform is applied to Eq. (20) a solution ψ of Eq. (5) will be expressed in the form

$$\psi(x,E) = \psi_0(x,E) + \int_{-\infty}^x K(x,y)\psi_0(y,E)dy \quad (25)$$

where $\psi_0(x, E)$ satisfies Eq. (5) when $P(x)$ is zero and $\psi(x, E) \sim \psi_0(x, E)$ as $x \rightarrow -\infty$. In Eq. (25) the independent variable r has been replaced by x to indicate that the range of the independent variable is $(-\infty, \infty)$ rather than $(0, \infty)$.

In particular, Eq. (25) holds when E is zero. Therefore, because of Eqs. (8) and (9),

$$Z^{-1/2} = Z_0^{-1/2} \left[1 + \int_{-\infty}^x K(x, y) dy \right], \quad (26)$$

which guarantees that for any given real, positive Z_0 $\lim_{x \rightarrow -\infty} Z = Z_0$.

To prescribe Z_1 it is necessary to choose an appropriate value of $r(0)$, the limiting static value of the reflection coefficient. Since this corresponds to the case in which the wavelength is infinite it might be expected, at least for a finite inhomogeneous transmission line, that $r(0)$ corresponds to the vanishing of the length of the inhomogeneous line. Thus, it would be expected that

$$r(0) = \frac{Z_1 - Z_0}{Z_1 + Z_0} \quad (27)$$

which is the usual value of the reflection coefficient at the interface between two homogeneous transmission lines having characteristic impedances Z_0 and Z_1 .

To see that Eq. (27) is, in fact, the correct condition consider that because of the properties (11), $r(0)$ and $t(0)$ are both real and that, because of Eq. (1), V is constant over the inhomogeneous line at zero frequency. Now from the definition of r and t and from Eq. (8),

$$VZ_0^{-1/2} = 1 + r(0) \quad VZ_1^{-1/2} = t(0) \quad (28)$$

It then follows from energy conservation (11) and from (28) that

$$Z_0 [1 + r(0)]^2 = t^2(0) Z_1 = [1 - r^2(0)] Z_1 \quad (29)$$

Equation (29) has two roots. One of these is $r(0) = -1$, which is ruled out by energy conservation except when the line is short circuited. The other root is given by Eq. (27).

A heuristic argument similar to the one just given leads to the fundamental integral equation of Marcenko. In this case the incident wave is from the right and therefore has the form $\delta(t + x)$, while the reflected wave has the form $S(t - x)$. The argument immediately leads to the integral equation for the corresponding kernel $H(x, t)$, which transforms a solution of Eq. (16) into a solution of Eq. (15):

$$0 = S(t - x) + H(x, -t) + \int_x^\infty H(x, y) S(t - y) dy \quad t < -x$$

If t is replaced by $-t$ the equation becomes

$$0 = S(-t - x) + H(x, t) + \int_x^\infty H(x, y) S(-t - y) dy \quad t > x \quad (30)$$

where

$$S(-\tau) = \frac{1}{2\pi} \int_{-\infty}^{\infty} \rho(\omega) e^{i\omega\tau} d\omega$$

In the Marcenko method the reflection coefficient ρ is not actually the one used. Instead, the boundary condition $\psi(0) = 0$ is imposed, and the corresponding reflection coefficient ρ_0 then has a modulus of 1 so that its Fourier transform will not exist as a proper function. This difficulty can be resolved by considering solutions $\psi_{\pm}(x)$ of Eq. (5) having the asymptotic forms

$$\psi_{\pm}(x) \sim \exp(\pm i\sqrt{E}x) \text{ as } x \rightarrow \infty$$

These solutions can be represented in accordance with Eq. (30), as

$$\psi_{\pm}(x) = \exp(\pm i\sqrt{E}x) + \int_x^\infty H(x, y) \exp(\pm i\sqrt{E}y) dy \quad (31)$$

The boundary condition will be satisfied by the solution $\psi_- + \rho_0 \psi_+$; hence,

$$\psi_-(0) + \rho_0(E) \psi_+(0) = 0$$

or

$$\rho_0(E) = \frac{-\psi_-(0)}{\psi_+(0)} = \frac{-[1 + \int_0^\infty H(0, y) \exp(-i\sqrt{E}y) dy]}{1 + \int_0^\infty H(0, y) \exp(i\sqrt{E}y) dy} \quad (32)$$

It follows that $1 + \rho_0(E) = 1 - \psi_-(0)/\psi_+(0) = o(1)$ for large real \sqrt{E} . The Fourier transform of $1 - \psi_-(0)/\psi_+(0)$ with respect to \sqrt{E} thus exists and gives the transform of ρ_0 except for a delta function. Since the incident disturbance does not reach $x = 0$ until $t = 0$, it must produce a reflected wave independent of the boundary condition at $x = 0$ until $t = 0$. It must therefore be equal to $S(\tau)$ for $\tau < 0$ or $S(-\tau)$ for $\tau > 0$. Thus, from Eq. (30)

$$0 = F(t+x) + H(x,y) + \int_x^\infty H(x,y)F(t+y)dy \quad (33)$$

where

$$F(\tau) = \frac{1}{2\pi} \int_{-\infty}^\infty \left[1 - \frac{\psi_-(0)}{\psi_+(0)} \right] \exp(i\sqrt{E}\tau) d\sqrt{E} \quad (34)$$

which is Marcenko's result.

Sharpe [1963] was able to adapt this solution of the inverse scattering problem for Eq. (5) to the transmission line problem despite the fact that it is based on a reflection coefficient corresponding to the boundary condition $\psi = 0$ at $x = 0$ by observing that the Fourier transform of the kernel F in Eq. (33) can be expressed directly in terms of elements of the scattering matrix for the transmission line without the short circuit at zero. If the line is assumed to be homogeneous for negative x , the solution ψ corresponding to a wave of unit amplitude incident from the left will be given by $\psi = r(\omega)\psi_+$ where the frequency ω now replaces \sqrt{E} of the quantum mechanics problem. For negative x

$$\psi = \exp(ik_0 x) + r(\omega)\exp(-ik_0 x)$$

Hence,

$$\psi_+(0) = \frac{1 + r(\omega)}{r(\omega)} \quad (35)$$

Also, from the properties (11) and Eq. (31) it follows that

$$\psi_-(0) = \frac{1 + r^*(\omega)}{r^*(\omega)} \quad (36)$$

Relations (35) and (36) permit the construction of the kernel F when $r(\omega)$ and $r^*(\omega)$ are known. Thus, the use of Marcenko's version of the Gelfand-Levitan technique for the transmission line problem, introducing the unnecessary boundary condition at zero, leads to a requirement for knowledge of two elements of the scattering matrix, while the direct approach leads to the solution in terms of Eq. (22), which requires a knowledge of just the reflection coefficient $r(\omega)$. As indicated earlier, $r(\omega)$ can be determined when $r(\omega)$ is known with the use of Eq. (12), and it is given explicitly by Eq. (13) when $r(\omega)$ is a rational function. In general, it would appear to be a cumbersome procedure, however.

APPROXIMATE SOLUTIONS

Here we review briefly some of the approximate analytical methods for the inverse scattering problem, but numerical methods as such are not considered. One of the greatest benefits to be derived from an approximate method is the possibility that it will provide some physical insight that might not be readily obtained from a more complicated exact method. Approximate methods are particularly useful if they shed some light on what is most important in the experimental data and what properties of the transmission line can be deduced from incomplete or partially erroneous data.

For many years, an approximate method has been used to solve the inverse scattering problem associated with vertical incidence radio soundings of the ionosphere [Rydbeck, 1942]. At a given carrier frequency a pulse is sent vertically upward, and the time delay of the return is observed experimentally as a function of the frequency. These data are then used to determine the variation of electron density as a function of height by a method which is based essentially on the W.K.B. approximation to the solution of Eq (5). The method has been limited mathematically to the case of a single turning point; that is, the assumption is made that $P(x)$ is a positive, monotonically increasing function. In the quantum mechanical interpretation this is equivalent to solving Abel's problem in classical mechanics: to determine the shape of a hill, given the time it takes for a particle to go up and return as a function of the particle's initial velocity at the foot of the hill [Courant and Hilbert, 1931, p. 134]. In fact, both problems lead to the Abel integral equation.

When $P(x)$ is a monotonically increasing function the reflection coefficient $r(\omega)$ has unit amplitude due to total reflection at some height where the plasma frequency is equal to the carrier frequency of the incident pulse; hence, the phase of $r(\omega)$ constitutes all of the information given by the experiment. In accordance with the W.K.B. method, the phase is approximately

$$\phi(k) = 2 \int_{x_0}^h [k^2 - P(y)]^{1/2} dy \quad (37)$$

where, for c the velocity of light in vacuum, $k = \omega/c$ is the wave number and h satisfies the equation

$$k^2 = P(h) \quad (38)$$

The group time delay of a pulse is then given by

$$\frac{1}{c} \frac{d\phi}{dk} = \frac{2k}{c} \int_{x_0}^h [k^2 - P(y)]^{-1/2} dy \quad (39)$$

The left side of Eq. (39) is determined from the experimental data, and Eq. (39) is the Abel integral equation for which the method of solution is well known [cf. Whittaker and Watson, 1948, p. 229].

Karp and Shmoy's [1955] considered the inverse scattering problem for Eq. (5) given the transit time of a pulse propagating through the medium between two fixed points. They also used the W.K.B. approximation and thus started with a relation similar to Eq. (39) except for the limits of integration.

The transit time is given by

$$\tau_{12}(k) = \frac{k}{c} \int_{x_1}^{x_2} [k^2 - P(y)]^{-1/2} dy \quad (40)$$

As indicated earlier, the transmission coefficient, and hence the transit time obtained from the phase of the transmission coefficient, is not sufficient to determine $P(x)$ uniquely. However, Karp and Shmoys point out that in this approximation P can be determined as a function of a variable ξ , which is defined by

$$\xi = \sum_i (x_{i+1} - x_i) \epsilon_i \quad (41)$$

where the sum is over all values x_i of x such that

$$P = P(x_i) \quad (42)$$

and where

$$\epsilon_i = \begin{cases} 1, & \text{if } P(y) < P \text{ for } x_i < y < x_{i+1} \\ 0, & \text{if } P(y) > P \text{ for } x_i < y < x_{i+1} \end{cases}$$

That is, for every possible value of P the x_i are roots of Eq. (42) and ξ is then determined by Eq. (41), thus defining $\xi(P)$ and, by inversion, $P(\xi)$. Equation (40) then becomes

$$\tau_{12}(k) = \frac{k}{c} \int_0^{x_2 - x_1} [k^2 - P(\xi)]^{-1/2} d\xi \quad (43)$$

Karp and Shmoys were then able to transform Eq. (43) into an integral equation that can be solved by means of a Mellin transform.

An approximate solution to the direct scattering problem for the transmission line equations (1), similar to the W.K.B. approximation used in connection with Eq. (5), leads more simply to a solution of the inverse scattering problem [Brunner, 1950]. For small $(dZ/dx)/Z$, Bremner [1951] has shown that a W.K.B. type of approximation is valid and is, in fact, the lowest term of a series that converges to the solution. For

$$F(z) = \frac{dZ/dx}{2Z} \quad (44)$$

the reflection coefficient, to its lowest nonvanishing order, can be approximated by [Kay, 1961]

$$r = \int_{x_0}^{\infty} G(x) \exp(-2ikx) dx \quad (45)$$

where $G(x) = F(z)$ when the change of variable (Eq. (8)) from z to x is introduced, and where

$$k = \omega(L_0 C_0)^{1/2}$$

The solution is obtained then, by taking a Fourier transform of both sides of Eq. (45) and integrating the result in accordance with Eq. (44).

Another method for solving the inverse scattering problem associated with Eq. (5) in the quantum mechanics case was given by *Jost and Kohn* [1952], and generalized by *Moses* [1956] to cover all scattering problems for Eq. (5), including higher dimensional cases and even cases where P is a linear operator of a very general type. The approximation is equivalent to the Neumann series expansion, of which the Born approximation is the lowest order term, for the solution of the direct scattering problem in a certain sense. If the series solution is cut off at a term of order n and the resulting approximation to P is used to calculate a Neumann series approximation of the direct scattering problem to the term of order n , the scattering matrix elements will be reproduced as given.

A particularly simple expression for $P(x)$ in the transmission line problem is obtained by this method when the first two terms of the expansion are used:

$$P(x) \sim -4[R'(2x) + R^2(2x)] \quad (46)$$

where R is defined by Eq. (18). This result is exactly what is obtained from the Gelfand-Levitan method if the integral equation (Eq. (22)) is solved by iteration to two orders.

SOME PROBLEM AREAS

In applying the various methods of solution to the practical determination of inhomogeneous transmission line characteristics, two questions have considerable importance: How sensitive will the result be to errors in the data, and what physical features can still be determined when the data are incomplete? In general, there are reasons to believe that the inverse scattering problem is very sensitive to certain kinds of error in the data. For example, the scattering matrix for Eq. (5) can be calculated approximately in the direct scattering problem by the variational method, the validity of which implies that small changes in the data could produce large changes in the wave function ψ . This may not be significant for the inverse scattering problem associated with Eq. (5) itself, but because of Eq. (9) the solution to the inverse scattering problem for the transmission line equations (Eq. (1)) is given in terms of a wave function of Eq. (5) with $E = 0$. Thus, it would not be surprising to find that the characteristic impedance has an unstable pointwise dependence on the reflection coefficient.

For inverse scattering problems associated with Eq. (5) there is no reason to expect that such an instability exists when $P(x)$ is a monotonically increasing function, a case for which the Abel integral equation approximation provides a reasonable result. The situation is not as clear, however, when $F(x)$ oscillates — represents a medium consisting of several layers as in the case of the real ionosphere. Shielding by front layers might completely obscure layers farther back. This can happen if, at the lower frequencies, the sounding pulse cannot penetrate far enough to affect the rear layers, and if the higher frequencies, for which sufficient penetration occurs, are above the critical frequencies of the rear layers.

Mathematically, these problems are connected with the fact that the kernel $R(y + x)$ of the integral equation (22) is the Fourier transform of the reflection coefficient and can be very sensitive to pointwise errors in the data. If the reflection coefficient is approximated by a rational or meromorphic function, the important properties of the data will be in the location of the poles and their corresponding residues. Although an explicit solution to the inverse scattering problem for Eq. (5) can be given when $r(\omega)$ is rational [Kay, 1960], the form of the general solution

is too complicated to exhibit in an obvious way the manner in which various layers of $P(x)$ might be correlated to specific features of the data. It would seem reasonable that poles far from the real axis will contain the information obscured by a shielding effect if it occurs. It is possible that this relationship can be studied and clarified by considering the reflection coefficient $\rho(\omega)$ on the far side. This is theoretically possible since $\rho(\omega)$ can be expressed in terms of $r(\omega)$ with the aid of relations such as Eqs. (11) and (12).

The question of how $r(\omega)$ can best be approximated by means of a rational function from given reflection measurements is important in any case since the solution to the inverse scattering problem can then be given explicitly. The approximation must be such that all known physical properties of the transmission line are taken into account—for example, that the line is finite in length and is terminated by an impedance equivalent to the characteristic impedance of a known homogeneous line of infinite length. *Heim and Sharpe* [1966] have discussed some such requirements of the transmission line and have pointed out that for a finite passive line $r(\omega)$ must be meromorphic with poles only in the lower half plane, it must be $o(1)$ in the upper half plane and on the real axis, and it must be $o[\exp(2ik_0 \ell)]$ in the lower half plane if the inhomogeneous line has length ℓ .

Another problem, which may not be important in the solution of a practical inverse scattering problem, is nevertheless disconcerting. The exact methods used for solving the inverse scattering problem associated with Eq. (1) cannot produce a line whose characteristic impedance has discontinuities if the mathematically required conditions of regularity are imposed [*Heim and Sharpe*, 1966]. For insight and for the sake of completeness it would be desirable to generalize the methods to include discontinuous solutions.

ADDENDA

A number of open problems exist in connection with the inverse scattering problem for transmission lines. Some of these were mentioned earlier. The following are added here primarily because they are probably solvable.

Two of the approximate methods for the inverse scattering problem associated with Eq. (5) have been derived from the W.K.B. approximation for the direct problem but have not been derived from the exact solution to the inverse scattering problem obtained by means of the integral Eq. (22). These are the Abel integral equation method for data in the form of the time delay of a reflected pulse, and the method of Karp and Shmoys for data in the form of the transit time of a pulse between points on either side of the inhomogeneous line. It would certainly be of academic interest, at least, to derive these approximations from the exact method. In addition, the derivation might lead to intermediate approximations that would improve on the original approximations or provide additional insight into the physical relationship between the data and the transmission line characteristics.

The restriction of the Abel integral equation method to cases in which $P(x)$ is monotonic increasing might be relaxed by such a generalization, thus providing a systematic, practical technique for interpreting existing ionospheric data. If the method of Karp and Shmoys were correlated to an exact treatment of the problem the precise nature of the information obtained in the transmission coefficient alone could be elucidated. This is related to the very practical question of what information can be obtained from partial scattering data.

Another open question is the solution of the inverse scattering problem for the $2n$ wire coupled system of transmission lines. *Jost and Newton* [1955] and *Agranovich and Marcenko* [1958] have given the solution for a system of coupled Schrödinger equations [cf. *Butler*, 1969, who considered the inverse scattering problem for differential equations of higher order than two]. A difficulty in applying Marcenko's method to the $2n$ wire coupled transmission line, as was done for a single 2 wire line, is the fact that, in general, the wave velocity $v(z)$ will be different for each channel and cannot, therefore, be dealt with by a simple change of variables such as that of Eq. (8). One possibility is to generalize or adapt the approximate method of Moses to the multiple transmission line problem. Some clue for handling the difficulty of different wave velocities might thereby be revealed so that the exact method of Marcenko might finally be applied.

DISCUSSION

Newton: The problem you mentioned at the end of coupled transmission lines is analogous to the quantum mechanical problem of coupled channels. It is analogous to coupled channels, inelastic scattering. Prof. Cox is the expert on that. You said there is a relation between the transmission coefficient and reflection coefficient. If you do the Marcenko procedure, you don't have to put that relation in – the two come in independently.

Kay: His expression for the kernel that goes into the Gelfand-Levitan equation is given directly in terms of the phase shift.

Newton: If you don't put in the relation, but put them in independently, what happens?

Kay: That's hard to answer. You will be violating one of the conditions on the scattering matrix elements themselves. Whatever potential you get will violate that condition. It would have some kind of pathological behavior.

Chadan: In the case of a simple, infinite line, what happens if you impose the boundary condition at the origin? You have two solutions, one of which vanishes at the origin. In matrix form you have the right to left and left to right signals.

Kay: This is what Marcenko's method does. You arrive at the boundary condition by taking a linear combination of the two fundamental waves. Then you get a relationship that involves both the transmission and reflection coefficients, which creates a problem because they are not independent.

Sabatier: You have spoken only of passive lines. Are there analogous results with nonpassive lines with sources?

Kay: I don't know of anyone who did that. When you have sources, the conditions imposed on the analytic behavior of the transmission coefficient are no longer applicable. This is analogous to what happens to the reliability conditions in circuit theory. For passive circuits, all your poles are confined to one half plane and they don't lie on the imaginary axis. For active circuits they can. You have to go back and look at the consequences, including the question of existence and uniqueness of the solution of the Gelfand-Levitan equation. In proving this, a very important contribution is the conservation of energy.

Mitra: You started with inductance and capacitance in the transmission lines and ended up with one function. That ambiguity can never be resolved.

Kay: You can assume anything you want for the other function as long as it is monotonic.

Mitra: If the transmission line had losses and you know the capacitance, resistance, and conductance, can the inductance be found?

Kay: No. They are independent.

THE INVERSE PROBLEM FOR RADIATION SCATTERING

Reese T. Prosser

Dartmouth College

ABSTRACT

N73-11613

This paper provides a brief survey of recent theoretical progress on the inverse scattering problem for radiation scattering from reflective boundaries and variable indices of refraction.

INTRODUCTION

The theory of radiation scattering is concerned primarily with the far-field relations between incident and scattered radiation in the presence of a scattering object. The primary, or direct, problem of this theory is to develop quantitative information about these scattering relations from a knowledge of the scattering object. The secondary, or inverse, problem, on the other hand, is to determine the nature of the object from an analysis of the scattering relations.

While the direct problem has by now generated a well-developed discipline, the inverse problem has only recently aroused any serious attention. This is especially surprising in view of the manifold practical applications to the design and operation of precision radar systems. This paper provides a brief survey of the present state of knowledge and possible future development of the inverse problem for radiation scattering.

SCATTERING FROM A POTENTIAL

The first real contribution to the inverse problem was made, not by radar engineers, but by nuclear physicists, around 1950. The physicists were interested in the analogous inverse problem for the scattering of electrons from a nuclear potential: to determine the shape of the potential from the resulting scattering data.

Levinson [1949] first showed that every spherically symmetric potential satisfying suitable regularity conditions, is uniquely determined by a knowledge of the phase shifts in the partial wave analysis of the scattering data. His proof however, gives no algorithm for constructing the potential from a given set of phase shifts. *Jost and Kohn* [1952] then provided a recursive algorithm for this purpose, which converges if the potential is sufficiently weak and there are no bound states. The recursion relations are clumsy to use in practice, however, and the conditions required on the potential are too restrictive for many applications. Meanwhile, *Gelfand and Levitan* [1951] showed that under suitable regularity conditions the potential may be recovered from the solution of a certain integral equation whose kernel depends on the spectral weights. These weights in turn may be obtained from a knowledge of the phase shifts and binding energies of the problem. Since then a number of variants and refinements of the Gelfand-Levitan procedure have appeared; they are summarized in the excellent survey of *Faddeev* [1963].

In this analysis it suffices to know the phase shifts and binding energies for any single value of the angular momentum. *Wheeler* [1955] and *Newton* [1967] have shown that a quite similar analysis reconstructs the potential from the phase shifts for all values of the angular momentum at any single value of the energy. In a sense, these results complement each other and exhaust the possibilities for radial potentials.

For potentials that are not spherically symmetric, the situation is more complicated in practice but essentially the same in principle. In a series of papers devoted to this problem, *Kay and Moses* [1955-1956] have shown that the Gelfand-Levitan procedure can be extended to nonspherically symmetric problems in such a way that the potential can be recovered from the solution of an integral equation whose kernel depends ultimately on the matrix elements of the scattering operator. For this procedure it suffices to know, at least in principle, the scattering data (cross sections and phase shifts) for a single aspect angle, all scattering angles over a fixed hemisphere, and all energies. The *a priori* conditions required of the potential, however, are not investigated in their papers.

Moses [1956] also has shown, by means of a similar extension of the Jost-Kohn procedure, that the general potential can be recovered, at least in principle, from a knowledge of the *backscattering* data for all aspect angles over a fixed hemisphere and all energies. Again the *a priori* conditions required are not clear, although the procedure appears to be applicable for sufficiently weak potentials.

None of these results applies directly to the inverse problem for radiation scattering, which involves scattering from a variable index of refraction, from a reflective boundary, or from both, rather than from a nuclear potential. In the first case we seek to determine the structure of the index of refraction rather than the structure of the potential, so that we are dealing with an inhomogeneous problem whose inhomogeneous term is frequency dependent. In the second case we seek to determine the shape of the boundary, so that we are dealing with a homogeneous boundary value problem rather than an inhomogeneous problem without boundary. Moreover, radiation scattering involves the vector wave equation rather than the scalar wave equation, although this difference does not appear to be decisive.

SCATTERING FROM A VARIABLE INDEX OF REFRACTION

Superficially, the problem of radiation scattering from a region of variable index of refraction strongly resembles the problem of electron scattering from a nuclear potential, and similar methods might be expected to apply. Specifically, the reduced wave equation for both problems has the form

$$\nabla^2 \phi + \omega^2 \phi = F(x, \omega) \phi$$

where in the case of potential scattering the inhomogeneous term

$$F(x, \omega) = V(x)$$

is independent of the frequency ω , while in the case of refraction scattering the inhomogeneous term

$$F(x, \omega) = \omega^2 [1 - n^2(x)]$$

is frequency dependent. For fixed frequency, then, the solutions to the two direct problems have the same form, but their dependence on frequency is different; and this difference forces a considerable difference in the solutions of the associated inverse problems. In particular, the Gelfand-Levitan procedure depends essentially on the frequency independence of the inhomogeneous term.

In the spherically symmetric case the index of refraction depends upon only the radial variable and the problem can be presented in one-dimensional terms. *Kay and Moses* [1955] and *Moses and DeRidder* [1963] have shown that in this case the refraction problem can be reduced to the potential problem by means of a form of Liouville transformation, and that the corresponding inverse problem can then be resolved by means of the Gelfand-Levitan procedure. Specifically; in this case the refraction problem reduces, via a partial wave analysis, to the one-dimensional problem:

$$\left(\frac{d^2}{dr^2} + \omega^2 n^2(r) \right) \phi(r) = 0$$

$$\phi(0) = 0 \quad \phi'(0) = 1$$

If we introduce a new independent variable

$$\rho(r) = \int_0^r n(r') dr'$$

and a corresponding new dependent variable

$$\psi(\rho) = n^{1/2}(r)\phi(r)$$

then we find that $\psi(\rho)$ satisfies the equation

$$\left(\frac{d^2}{d\rho^2} + \omega^2 \right) \psi(\rho) = V(\rho)\psi(\rho)$$

where the "potential" $V(\rho)$ has the form

$$V(\rho) = n^{-1/2} \frac{d^2}{d\rho^2} \left(n^{1/2} \right)$$

$$V(\rho) \text{ and } V'(\rho) \rightarrow 0 \text{ as } \rho \rightarrow \infty$$

Now V , and hence n , may be deduced from a knowledge of the S-wave phase shifts via the Gelfand-Levitan algorithm. Note that the new independent variable here is essentially the optical path length, and that, in the new variables, the optical path metric is constant.

In principle, this procedure can be extended to nonspherically symmetric problems, but in practice new difficulties intrude. If we start from the reduced wave equation

$$[\nabla^2 + \omega^2 n^2(\mathbf{x})] \phi(\mathbf{x}) = 0$$

and introduce the Ansatz

$$\phi(\mathbf{x}) = e^{ikS(\mathbf{x})}\psi(\mathbf{x})$$

then we find

$$(\nabla^2 + \omega^2 n^2) \phi = e^{ikS} \psi [\omega^2 (n^2 - |\nabla S|^2) + i\omega (\nabla^2 S + 2\psi^{-1} \nabla \psi \cdot \nabla S) + \psi^{-1} \nabla^2 \psi]$$

Hence, if we choose S so that

$$|\nabla S|^2 - n^2 = 0$$

$$kS(x) \rightarrow k \cdot x \text{ as } |x| \rightarrow \infty$$

and then choose $\psi_0(x)$ so that

$$2\nabla \psi_0 \cdot \nabla S + \psi_0 \nabla^2 S = 0$$

then we find that, with $\phi_0 = \psi_0 e^{ikS}$,

$$(\nabla^2 + \omega^2 n^2 - \psi_0^{-1} \nabla^2 \psi_0) \phi_0 = 0$$

Hence, the solution of the original equation must also satisfy an equation of the form

$$(\nabla^2 + \omega^2 n^2 - \psi_0^{-1} \nabla^2 \psi_0) \phi = (-\psi^{-1} \nabla^2 \psi_0) \phi$$

The right-hand side here is independent of ω and serves as a potential, for which a suitable extension of the Gelfand-Levitan procedure would apply.

Note that here S satisfies the eikonal equation, and ψ_0 is related to the gaussian curvature of the level surfaces of S . In effect, we have introduced as new coordinates essentially the rays and wave fronts of the geometric optics solution, which is frequency independent. This transformation is nonsingular, however, only if there are no caustics in the geometric optics solution. The presence of caustics presents additional difficulties which have not yet been resolved.

The Gelfand-Levitan procedure is not even formally applicable to problems involving frequency-dependent perturbations. The Jost-Kohn procedure, on the other hand, is not sensitive to such a frequency dependence and normally extends directly to the general refraction problem. Briefly, this procedure begins by expanding the reactance matrix T in the Born series [Prosser, 1969]:

$$T(k', k) = k^2 w(k' - k) + \int k^2 w(k' - k'') \frac{k^2}{k''^2 - k^2} w(k'' - k) dk'' + \dots$$

where k and k' are the incoming and outgoing momenta, $k^2 = k'^2 = \omega^2$, and $w(k)$ is the Fourier transform of $1 - n^2(x)$. The reflection, or backscatter, coefficient $b(k)$ is simply the value of $T(k', k)$ at $k' = -k$. If we now replace $b(k)$ by $\epsilon b(k)$, and $w(k)$ by $w_\epsilon(k) = \sum_{n=1}^{\infty} \epsilon^n w_n(k)$, substitute in the expansion above for $T(-k, k)$, and equate the coefficients of ϵ^n , we can solve for the $w_n(k)$ in terms of the reflection coefficient $b(k)$:

$$\begin{aligned}
k^2 w_1(k) &= b(k) \\
k^2 w_2(k) &= \int k^2 w_1(k' - k'') \frac{k^2}{k''^2 - k^2} w_1(k'' - k) dk \\
&\dots \\
k^2 w(k) &= \sum k^2 w_n(k)
\end{aligned}$$

This procedure determines the index of refraction, at least in principle, from a knowledge of the backscatter coefficient for all aspect angles and all frequencies. A similar procedure determines the index of refraction from the scattering angles and all frequencies, or for fixed scattering angle, all aspect angles and all frequencies, or for certain other combinations of data. These solutions are all formal, and give no information on the question of convergence or the *a priori* restrictions required of the index of refraction. Moses [1956] has shown, however, that in the spherically symmetric case the first two approximations to the index of refraction obtained by this method agree with the first two approximations obtained by the Gelfand-Levitan procedure, so that there is some reason to be optimistic about the prospects.

SCATTERING FROM A REFLECTIVE BOUNDARY

For the problems of radiative scattering from reflective boundaries, the preceding analysis does not directly apply. Certain approximate results are available, however, which suggest various avenues for future study.

If the boundary is known to be smooth and convex, with a gaussian curvature everywhere positive, and if the frequencies involved are sufficiently high, then the geometric optics approximation to the scattering matrix may be valid. According to this approximation, the scattering cross section measured through a scattering angle θ is given by

$$\sigma(\theta) = \frac{R\left(\frac{\theta}{2}\right)}{4\pi G\left(\frac{\theta}{2}\right)}$$

where $R(\theta/2)$ is the reflectivity coefficient at the reflection point on the surface (whose normal lies at an angle $\theta/2$ with the incident ray) and $G(\theta/2)$ is the gaussian curvature of the surface of that point [Keller, 1959]. Thus, if the reflectivity is constant over the surface, the scattering cross section is inversely proportional to the gaussian curvature at the reflection point. It follows that a knowledge of the scattering cross section for a single (high) frequency at two (diametrically opposite) aspect angles and all scattering angles, or alternatively, a knowledge of the backscattering cross section for a single frequency at all aspect angles, will determine the gaussian curvature of each point of the surface in terms of the direction cosines of the normal at the point. Note that the convexity of the surface assures that there is just one such point for each normal direction.

Thus, we are led to the problem of determining a smooth convex surface from a knowledge of its gaussian curvature. This problem is a classic in the geometry of surfaces, and goes by the name of Minkowski's problem. It is known that Minkowski's problem admits a unique solution provided that the gaussian curvature $G(n)$, as a function of the normal direction n , satisfies the obvious necessary condition:

$$\int_S G(n) dS = 0$$

If this condition is satisfied, the surface may be obtained as the solution of a certain nonlinear elliptic differential equation of Monge-Ampere type. [Nirenberg, 1953]. If the surface is described in rectangular coordinates in the form $z = \theta(x, y)$ then the determining equation has the form

$$\frac{rt - s^2}{(1 + p^2 + q^2)^2} = G(p, q, 1)$$

where $p = \partial\theta/\partial x$, $q = \partial\theta/\partial y$, $r = \partial^2\theta/\partial x^2$, $s = \partial^2\theta/\partial x\partial y$, $t = \partial^2\theta/\partial y^2$, and $G(p, q, 1)$ denotes the gaussian curvature of the point whose normal has direction numbers $(p, q, 1)$.

The proof of the existence and uniqueness of the solution, however, is nonconstructive, and no effective algorithm for constructing the solution has appeared.

Special cases of this result are sometimes useful. If, for example, the surface is known a priori to be *axially symmetric*, it may be described in the form $x = \theta(\rho)$ where $\rho^2 = y^2 + z^2$. Then the determining equation becomes

$$\frac{\phi' \phi''}{\rho(1 + \phi'^2)^2} = G(\phi')$$

where now $G(\phi')$ denotes the gaussian curvature given as a function of the normal direction $(1, \theta' \cos \alpha, \theta' \sin \alpha)$. the axial symmetry implies that the gaussian curvature is the same for all values of $\alpha = \arctan z/y$ and so depends only on the slope θ' . Hence, it suffices to know the curvature only at points with different values of θ' —that is, only along any profile. In this case the determining equation is integrable for θ' , and hence for θ , by quadrature.

We conclude that for axially symmetric smooth convex bodies, it suffices to know the monostatic, monochromatic scattering cross section for each aspect in any plane containing the axis of symmetry.

If polychromatic scattering data are available, then we might expect that fewer aspect angles would be required. To see how this can come about, we consider again a smooth convex body, but replace the high-frequency incident wave with a short pulse and use the physical optics approximation in place of the geometric optics approximation. Then in this approximation the backscatter return from a short pulse incident on the surface is essentially proportional to the second derivative of the cross-sectional area of the surface sectioned at a distance ct from the source and normal to the incident direction. Thus we have $\sigma(2t) = \text{const } A''(ct)$ where $A(x)$ is the cross sectional area of the surface in a vertical plane normal to the x axis and at distance x from the origin, and $A''(x) = d^2A/dx^2$. Here the source lies at the origin, with the incident direction along the x axis [Lewis, 1969].

From these relations we can recover the cross-sectional area $A(x)$ as a function of x . If we now do the same for all other *horizontal* aspects, we obtain a two-parameter family of cross-sectional areas of the surface.

The problem of recovering a surface from a knowledge of its cross-sectional areas is another classic in surface geometry, and is known as Radon's problem. This problem, too, has a unique solution for smooth convex bodies

provided only that certain necessary conditions are satisfied. The solution may be expressed in terms of quadratures over derivatives of the cross-sectional area data. If, for example, the surface is described in rectangular coordinates as $z = \phi(x, y)$ and we are given the areas

$$A(r, \phi) = \int_{L(r, \phi)} \phi(x, y) dL$$

where $L(r, \phi)$ is the line in the xy plane given by $x \sin \theta + y \cos \theta = r$, then the function $\phi(x, y)$ may be recovered through the formula

$$\phi(x, y) = \int_0^{2\pi} d\theta \int_0^{A_{\max}} \frac{dA(r, \theta)}{r - p}$$

where p denotes the projection of the vector (r, θ) on the vector (x, y) .

Thus it suffices, at least in principle, to know the backscattering amplitude as a function of time resulting from a short pulse incident on the surface from each aspect in any plane passing through the surface.

If it is known a priori that the surface is axially symmetric, then it suffices to probe the surface with a short pulse in two diametrically opposite aspects. If these aspects are taken as "nose on" and "tail-on" along the axis of symmetry, the formulas above reduce to the following:

$$x = \phi(\rho) \quad \rho = (y^2 + z^2)^{1/2} = \psi(x)$$

$$\sigma(2t) = \text{const } A''(ct)$$

$$A(x) = \pi \rho^2 = \pi \psi(x)^2$$

It follows that

$$\psi(x) = [\pi^{-1} A(x)]^{1/2}$$

Similar formulas for other aspects may also be derived along the same lines.

The weakness in all of these arguments, of course, lies in the optics approximations, which are notoriously bad, particularly near the shadow line. Nevertheless these arguments make plausible the hypothesis that monostatic scattering data depending essentially on two parameters, such as two independent aspect angles or one aspect angle and the frequency, should suffice to determine the shape of a reflective boundary.

It would be highly desirable to have a procedure for reconstructing the shape of the boundary from the scattering data that, in principle at least, is exact. The analysis of *Lax and Phillips* [1963] has shown that every smooth closed reflective surface is uniquely determined by the scattering matrix — by the *totality* of the scattering data — so that the

search for a constructive algorithm is at least not doomed from the outset. Prosser [1969] has shown that, at least formally, the Jost-Kohn procedure can be extended to this problem as well, and that it reconstructs the Fourier transform of the characteristic function of the region bounded by the surface as a formal series of approximations involving the backscatter coefficient, just as for the general refractive problem. This procedure requires the backscatter data of all aspects and all energies, so that, in light of the preceding arguments, the solution is overdetermined. Moreover, in simple examples the first few terms of the approximation don't seem to give anything sensible at all. At the moment, however, this procedure is all we have.

One other approach deserves mention here. If we consider the boundary scattering problem in two dimensions, it can be reduced to a refraction scattering problem by means of a suitable conformal mapping. Specifically, suppose we are dealing with the problem

$$\nabla^2 \phi + \omega^2 \phi = 0 \quad \text{outside } S$$

$$\phi \text{ prescribed} \quad \text{on } S$$

defined in the xy plane. If we put $z = x + iy$, and introduce the conformal mapping $z = F(w)$ of the exterior of the unit circle C in the $w = u + iv$ plane onto the exterior of the given boundary S in the $z = x + iy$ plane, normalized so that $F(\infty) = \infty$, and $F'(\infty) = 1$, then the equation

$$\nabla_z^2 \phi + \omega^2 \phi = 0 \quad \text{outside } S$$

$$\phi \text{ prescribed} \quad \text{on } S$$

is mapped into the equation

$$\nabla_w^2 \phi + \omega^2 n^2(w) \phi = 0 \quad \text{outside } C$$

$$\phi \text{ prescribed} \quad \text{on } C$$

where the "index of refraction" is just $n(w) = |F'(w)|$. Because of the normalization of $F(w)$, $n(w) \rightarrow 1$ as $|w| \rightarrow \infty$ and the scattering data are all unchanged. Thus, any procedure that resolves the inverse refraction problem will reconstruct $n(w) = |F'(w)|$ from the scattering data, and since the mapping is conformal, $F(w)$ and hence S can be reconstructed from $|F'(w)|$.

In two dimensions, then, the boundary problem becomes a special case of the refraction problem, and will yield to the same methods. In three dimensions, however, the method of conformal mapping is not available, and no similar reduction is yet known.

PRACTICAL CONSIDERATIONS

Even if exact theoretical solutions were available for the general inverse problems of refractive and boundary radiative scattering, the task of translating these solutions into practical procedures for the design and operation of precision radar systems would still present many problems.

Let us suppose, for example, that such a system is to be used to identify the structure of an unknown object in orbit above the earth. Then to begin with, the index of refraction along the radar sight path is generally not constant, and will depend on meteorological and other natural data as well as the frequency and other operating parameters. In addition, the general shape and reflectivity of the unknown object may not be available. In particular, the surface certainly need not be simple, smooth or convex, and the reflectivity need not be constant over the surface. Moreover, the aspect angles are not under the control of the system operator. It is true that most objects in orbit contain some mechanical lossy mechanism, so that their spin axis eventually converts to the axis of largest moment of inertia and they tumble end-over-end; it is also true that such an object eventually presents every aspect to a fixed radar installation. Nevertheless, the task of extracting the scattering data for each aspect of an object whose motion is unknown from a time record of observations made over many passes, is by no means trivial. Finally, if the object makes only one pass, most aspects are never seen at all.

Practically speaking, the scattering data obtained in any such application are apt to be inexact and incomplete, but the theoretical procedures described above for recovering the object from the scattering data all assume a complete and exact knowledge of this data. No studies have been made of the stability question: whether small errors in the data cause only small errors in the result. And no progress has been reported on the approximation question: what partial data can be made to yield partial results. It might be possible, for example, to infer from suitable partial data the first few spherical moments, or other similar signature parameters, of the unknown object. At present we have no information in this direction.

Finally, even if all of these problems could somehow be resolved, the remaining task of computing the shape of the object from the scattering data is in general nontrivial. If, for example, the object is known to be smooth and convex, and the exact geometric optics backscatter data are obtained from all aspects, the problem remains of computing the shape of the surface from a knowledge of these data, even though the solution is known to exist. The trouble is that the required shape is obtained as the solution of a nonlinear elliptic equation for which the usual computing algorithms fail miserably. The errors appear to accumulate uncontrollably if local approximations are used. In spite of considerable work on this problem (mostly classified) no successful computational procedure has been reported. [GISAT, 1965]

DISCUSSION

Newton: I'm a little puzzled by your statement that you can generalize the Jost-Kohn procedure to three dimensions. The first step is to equate the scattering amplitude to the first Born approximation. The Born term depends only on the difference between k and k' , but the scattering amplitude will not. It will depend on both vectors independently. I'm not sure how to start.

Prosser: The first approximation for the potential, in the case of potential scattering, is simply given by
$$V_1(2k) = T(-k, k).$$

Chadan: When you expand the T matrix as a Born series, it doesn't converge if you have bound states. How do you know if you have bound states or not?

Prosser: Characteristically, the radiation problems have no bound states. But the convergence question is still open.

Chadan: When you go to higher terms, those integrals are principal value integrals.

Prosser: Yes, but one treats them with the usual quantum mechanical prescription, by assigning to the denominator a small positive imaginary part.

Bojarski: The geometrical optics problem and the Minkowski problem have been solved successfully by Keller and Lewis for a body of revolution [Keller 1959].

Kay: Do you think that the Jost-Kohn method might be related to a multipolar expansion for the body?

Prosser: It looks to me a little more like this functional integration procedure we heard about here.

Unidentified speaker: Since mention has been made of the papers of Faddeev, and since they represent considerable progress in the inverse problem they should be referenced here: Doklady May, September 1966. Presumably Faddeev has new results.

Prosser: The suggestion is that a combination of his results and the methods of Kay and Moses might well make the Gelfand-Levitan algorithm accessible in three dimensions. That would be a tremendous advance.

DIAGNOSTICS OF NONRADIALLY STRATIFIED MEDIA

J. Shmoys and J. Pirraglia

Polytechnic Institute of Brooklyn

INTRODUCTION

1 N73-11614

With a few exceptions, the inverse scattering methods have been applied to one-dimensional distributions only. Both the trajectory approaches of the Abel and Herglotz type and the wave approaches of the Gelfand-Levitan type were directed to plane, cylindrically, and spherically stratified distributions.

In recent years a number of papers [Moses, 1956; Kay and Moses, 1961; Prosser, 1969] dealt with the application of the wave-type methods to asymmetric distributions. The difficulties with this approach are twofold. First, the wave methods are much more complicated than trajectory methods. Second, the Jost and Kohn approach used by Moses and Prosser is based on a Neumann (Born) series, which converges slowly if the distribution to be diagnosed is very extensive and of sufficient magnitude to produce large phase shifts.

Here we investigate an admittedly primitive method, of the trajectory type, to distributions lacking spherical (ϕ , in two dimensions, cylindrical) symmetry. We assume that the medium is "thin" enough so that the ray remains rectilinear. The data to be used consists of the change in phase of a wave propagated along a ray due to the presence of the medium, for a given set of rays.

TWO-DIMENSIONAL DISTRIBUTION

Let us consider the problem of diagnosing a two-dimensional refractive index distribution $n(r, \theta)$. The set of probing rays includes all straight lines. The parameters characterizing a particular ray are the impact parameter b and ray direction α (fig. 1). The refractivity of the medium will be assumed expandable in a Fourier series:

$$n(r, \theta) - 1 = \sum_{-\infty}^{\infty} N_k(r) e^{ik\theta} \quad (1)$$

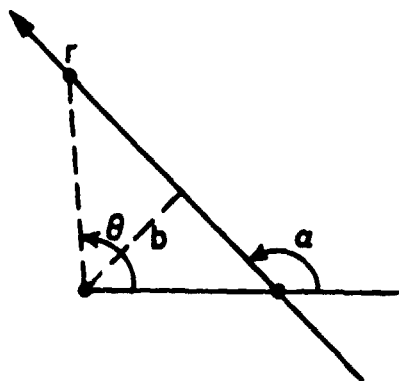


Figure 1.— Ray configuration

It is difficult to conceive of a physical situation in which the representation (fig. 1) would not exist. The phase shift for a ray is then

$$\begin{aligned}
\Phi(b, \alpha) &= \frac{\omega}{c} \int_{\alpha-\pi}^{\alpha} \sum_k N_k \frac{b}{\sin(\alpha-\theta)} e^{ik\theta} \frac{b d\theta}{\sin^2(\alpha-\theta)} \\
&= \frac{\omega}{c} \sum_k \int_0^{\pi} N_k \frac{b}{\sin \beta} e^{ik(\alpha+\beta)} \frac{b d\beta}{\sin^2 \beta} \\
&= \frac{\omega}{c} \sum_k e^{ik\alpha} \int_0^{\pi} b N_k \frac{b}{\sin \beta} e^{ik\beta} (\sin \beta)^{-2} d\beta \\
&= \frac{\omega}{c} \sum_k e^{ik\alpha} I_k(b)
\end{aligned} \tag{2}$$

where

$$I_k(b) = \int_0^{\pi} b N_k \frac{b}{\sin \beta} e^{ik\beta} (\sin \beta)^{-2} d\beta \tag{3}$$

We can simplify this further by recognizing that the integrand in Eq. (3) has symmetry about $\pi/2$:

$$N_k \frac{b}{\sin(\pi-\beta)} e^{ik(\pi-\beta)} \left[\sin(\pi-\beta) \right]^{-2} = e^{ik\pi} N_k \frac{b}{\sin \beta} e^{-ik\beta} (\sin \beta)^{-2}$$

so that

$$I_k(b) = \int_0^{\pi/2} b N_k \frac{b}{\sin \beta} \left[e^{ik\beta} + (-1)^k e^{-ik\beta} \right] (\sin \beta)^{-2} d\beta \tag{4}$$

The term in brackets reduces to $2 \cos k\beta$ for k even, and to $2i \sin k\beta$ for k odd. Hence I_k is either pure real or pure imaginary. The functions $I_k(b)$ are obtainable directly from the observed phase shifts, since they are

simply the coefficients of a Fourier series expansion of $\Phi(b, \alpha)$:

$$I_k(b) = \frac{c}{2\pi\omega} \int_0^{2\pi} \Phi(b, \alpha) e^{-ik\alpha} d\alpha \quad (5)$$

Equations (4) represent then a set of independent integral equations for the coefficient functions $N_k(r)$.

SOLUTION OF THE INTEGRAL EQUATION

It is easy to see that for $k = 0$, Eq. (4) is equivalent to the usual Abel equation for a radially stratified medium. Similarly, for $k = 1$, the integral equation is

$$\frac{1}{2i} I_1(b) = \int_0^{\pi/2} b N_1\left(\frac{b}{\sin \beta}\right) \sin \beta (\sin \beta)^{-2} d\beta$$

or

$$\frac{1}{2i} \frac{I_1(b)}{b} = \int_0^{\pi/2} b \left[\left(\frac{\sin \beta}{b} \right) N_1\left(\frac{b}{\sin \beta}\right) \right] (\sin \beta)^{-2} d\beta \quad (6)$$

so that we obtain an Abel equation for $N(r)/r$. Unfortunately, simple reduction to an Abel equation is not possible for $k > 1$. We have not been able to invert the equation explicitly. However, integral Eqs. (4) can be inverted numerically. To assess the difficulty of doing so, we tested a procedure that could be used for all k on special cases:

1. Select a set of values of b for which the data will be matched: b_0, b_1, b_2, \dots , in decreasing order.
2. For $r > b_0$ assume $N(r) = N_0 \exp [(b_0 - r)/H]$, when H is chosen arbitrarily.
3. Using the exponential approximation in the integral equation, evaluate N_0 .
4. For $b_0 < r < b_1$, assume $N(r) = Ar + B$.
5. From the requirements that (a) the function $N(r)$ be continuous at b_0 and (b) the linear-exponential approximation satisfy the integral equation at $r = b_1$, obtain two equations for the two unknowns A and B , that can be solved.
6. Proceed in a similar way with the piecewise linear approximation to the smallest value of b for which data are available.

We have inverted the phase shifts calculated for $N_0(r) = e^{-r^2}$ and $N_1(r) = e^{-1.2r^2}$ using the points 2, 1.8, 1.6, 1.4, 1.2, 1. Three values of H were used in each case: 0.5, 1, and 2. The results are shown of figures 2 and 3. We see that in spite of the crude piecewise linear model and sparsity of points, accuracy is fair. The choice of H has little effect beyond the initial point. The method is simple enough so that no elaborate programming was necessary. Clearly a denser set of points and/or the use of a piecewise parabolic model with continuous derivatives would have improved the results.

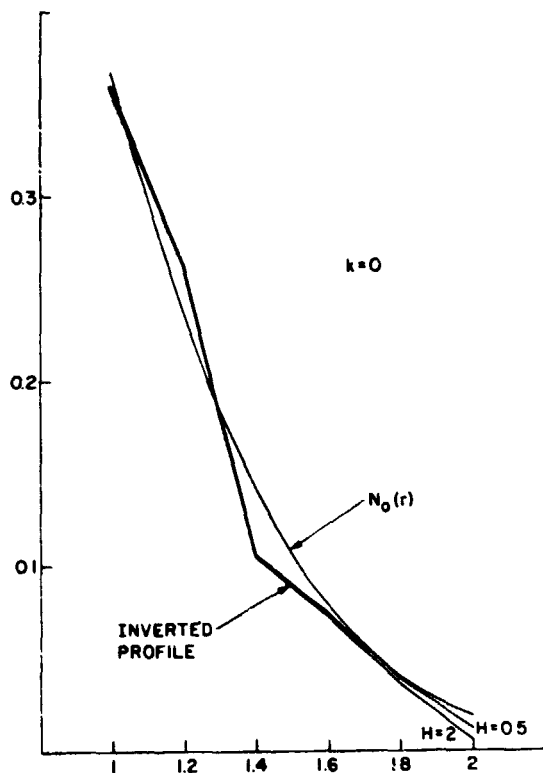


Figure 2. - Results of numerical inversion for $k = 0$.

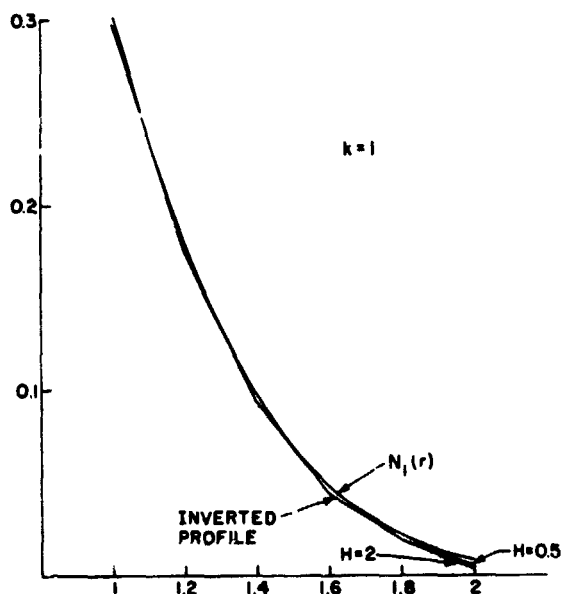


Figure 3. - Results of numerical inversion for $k = 1$.

Another procedure that could be used relies on the expansion of $N(r)$ in inverse powers of r :

$$N(r) = \sum_j A_{k,j} r^{-j} \quad (7)$$

so that

$$I_k(b) = \sum_j \left\{ A_{k,j} \int_0^{\pi/2} \left[e^{ik\beta} \Psi(-1) e^{-ik\beta} \right] (\sin \beta)^{j-2} d\beta \right\} b^{1-j}$$

If $I_k(b)$ can be represented as a series in inverse powers of b , the coefficients $A_{k,j}$ can be obtained. Such an expansion could be obtained by observing that for large b the lowest value of j dominates the behavior of $I(b)$; after that term is subtracted, the next higher value dominates, etc.

EXTENSION TO THREE-DIMENSIONAL DISTRIBUTIONS - APPLICATIONS

The method described above is capable of inverting a two-dimensional distribution in a given plane. Since we are assuming that rays remain linear, a three-dimensional distribution can be inverted by applying the two-dimensional procedure in a series of parallel planes (or any other convenient set of planes covering the region of interest sufficiently densely).

The methods outlined here could be used for diagnostics of planetary atmospheres provided sufficient experimental data were available. Note that for a three-dimensional distribution we would need a three-parameter set of data, whereas even a single-parameter family is hard to obtain. More realistically, perhaps, the method could be applied to millimeter wave (or laser) plasma diagnostics where we often encounter rotating cylindrical (but not cylindrically symmetric) distributions. The apparent rotation is caused by an azimuthally traveling instability.

The two-parameter set of data would be easier to obtain in this case since the angular parameter α would be equivalent to time. This substitution is possible because of the normally rather slow rate of rotation of the plasma distribution. Hence one would need only a set of phase shift versus time curves for different values of the impact parameter.

ACKNOWLEDGEMENTS

Research reported herein was sponsored in part by the Atomic Energy Commission under Grant No. AT(30-1)3956 and in part by the National Aeronautics and Space Administration under Grant No. NGR33-006-047.

DISCUSSION

Kay: You mentioned that the medium has to be weak. It seems that the approximation you used would be valid even if it isn't weak as long as it's slowly varying.

Shmoys: No. The approximation made here is that the rays are straight.

ELECTROMAGNETIC INVERSE SCATTERING*

Norbert N. Bojarski[†]

ABSTRACT

N73-11615

A three-dimensional electromagnetic inverse scattering identity, based on the physical optics approximation, is developed for the monostatic scattered far field cross section of perfect conductors. Uniqueness of this inverse identity is proven. This identity requires complete scattering information for all frequencies and aspect angles. A nonsingular integral equation is developed for the arbitrary case of incomplete frequency and/or aspect angle scattering information. A general closed-form solution to this integral equation is developed, which yields the shape of the scatterer from such incomplete information. A specific practical radar solution is presented. The resolution of this solution is developed, yielding short-pulse target resolution radar system parameter equations. The special cases of two- and one-dimensional inverse scattering and the special case of a priori knowledge of scatterer symmetry are treated in some detail. The merits of this solution over the conventional radar imaging technique are discussed.

INTRODUCTION

The direct scattering problem, whether electromagnetic, acoustic, particle, or quantum mechanical, is defined as the problem of predicting the scattered quantities, given the incident quantities, the relevant description of the scatterer, and the appropriate laws governing the interaction. The backscattering and forward scattering problems are the special monostatic case and the special bistatic case of scattering in the direction of incidence, respectively.

The inverse scattering problem is defined as the problem of determining the relevant quantities describing the scatterer, given the incident and scattered quantities, and the appropriate laws governing the interaction. The uniqueness and well-behavedness problems must also be taken as an integral part of the inverse scattering problem. The well-behavedness problem is the problem of determining the degree of smoothness and continuity with which the so-called "output data" vary with respect to the so-called "input data." The uniqueness problem must also deal with the question of incomplete input data.

The electromagnetic inverse scattering problem is thus defined as the problem of determining the size, shape and electromagnetic properties distributions (conductivity, susceptibility, and permeability distributions) of a scatterer, given the incident and scattered electromagnetic fields, and the electromagnetic field equations (Maxwell's equations and the appropriate wave equations; the constitutive equations and their coefficients are taken as part of the electromagnetic properties of the scatterer); and the determination of whether this problem is uniquely solvable for incomplete input data — the various permutations of incomplete bistatic aspect angles, incomplete monostatic aspect angles, incomplete frequency, monochromatic data only, incomplete polarization matrix, amplitude (power cross-section) data only, and scattered far field data only. Also of interest is the inverse scattering problem for which there is some a priori information about the scatterer, such as that the conductivity of the scatterer is infinite (perfect scatterer) or that the geometry of the scatterer is of axial symmetry.

This paper is restricted to the monostatic far field special case, a restriction [being characteristically] inherent to the radar application. For mathematical reasons that will become evident, the further special case of a priori

*This research was supported in part by USAF contracts AF 30 (602) 3961 and F 30602-68-C-0260.

[†]Physicist; Research Contractor and Consultant to the *Department of Defense*

knowledge of the scatterer being a perfect conductor is treated. For both these cases, special attention is given to the further special case of the finite three-dimensional aperture, the case for which monostatic scattering data are available for a limited and incomplete frequency and aspect angle domain. From a practical short-pulse target resolution radar point of view, this special case is of fundamental and primary interest.

THE PHYSICAL OPTICS CROSS SECTION

Here the physical optics approximation is taken as the basis for the direct scattering theory, and its validity for, and consistency with, short-pulse target resolution radar concepts is discussed briefly.

The scattered magnetic field H^s in terms of the induced (by the incident field) surface current density K on the surface of a perfect conductor is given by [Stratton, 1941]:

$$H^s = \oint_s \nabla \phi \times K \, ds \quad (1)$$

where the Green's function ϕ and its gradient are given by

$$\phi = \frac{e^{ikr}}{4\pi r} \quad (2)$$

$$\nabla \phi = \frac{ikr - 1}{r^2} \hat{r} \phi \quad (3)$$

The physical optics approximation [Van Bladel, 1964] for the induced surface current density in terms of the incident magnetic field H^i is

$$K = \begin{cases} 2n \times H^i, & \text{on the "illuminated" segment of } s \\ 0, & \text{on the "shadow" segment of } s \end{cases} \quad (4)$$

Thus, by Eqs. (1), (3), and (4), the scattered magnetic far field H^f , in terms of the wave number propagation vector k^s of the scattered far field, is

$$H^f = -2ik^s \times \int_{k^i \cdot n < 0} \phi n \times H^i \, ds \quad (5)$$

If the incident field H^i is taken as a plane wave of the form

$$H^i = I e^{ik^i \cdot x} \quad (6)$$

in the vicinity of the scatterer, where k^i is the wave number propagation vector of the incident field, and the range

and phase normalized (in the coordinate system in which the scatterer is described) scattered far field is taken as

$$H^f = \frac{1}{\sqrt{4\pi r^2}} S e^{ik^s \cdot x} \quad (7)$$

then for the monostatic case, that is,

$$k^s = -k^i \equiv k \quad (8)$$

Eq. (5) reduces with the aid of Eqs. (2) and (8), and the transversality of the incident field to

$$S = \frac{i}{\sqrt{\pi}} \int_{k \cdot n > 0} e^{-2ik \cdot x} k \cdot ds \quad (9)$$

Consistent with the conventional definition [Berkowitz, 1965] of the power cross section σ and the field cross section ρ

$$S = \rho I \quad (10)$$

$$\sigma = \rho \rho^* \equiv 4\pi r^2 \frac{|H^f|^2}{|H^i|^2} \quad (11)$$

Eq. (9) yields for the physical optics field cross section the well-known expression [Berkowitz, 1965]

$$\rho = \frac{i}{\sqrt{\pi}} \int_{k \cdot n > 0} e^{-2ik \cdot x} k \cdot ds \quad (12)$$

For a short-pulse target resolution radar system to be effective, its pulse length must be short compared to the target size; furthermore, since the fractional bandwidth of such a pulse is limited by practical considerations to much less than unity, it follows that the largest wavelength in the spectrum of the transmitted pulse must be very short indeed compared to the target size. The physical optics approximation Eq. (12) is thus a good model for the direct scattering theory for such a short-pulse target resolution radar system. For a detailed discussion of the physical meaning and implications of the physical optics approximation, that is, its being a total first order local scattering theory, consistent with short-pulse radar concepts, see Bojarski [1968].

THE INVERSE SCATTERING IDENTITY

With the aid of the physical optics approximation, we now develop a basic inverse scattering identity that states that the characteristic function (in three-dimensional space) of a scatterer is related to the field cross-section (in three-dimensional k-space) by a three-dimensional Fourier transform. Uniqueness of the solution for finite-sized scatterers is established.

Introducing the variable κ defined as

$$\kappa \equiv 2k \quad (13)$$

yields for Eq. (12)

$$\rho(\kappa) = \frac{i}{\sqrt{4\pi}} \int_{\kappa \cdot n > 0} e^{-i\kappa \cdot x} \kappa \cdot ds \quad (14)$$

Thus

$$\rho^*(-\kappa) = \frac{i}{\sqrt{4\pi}} \int_{\kappa \cdot n < 0} e^{-i\kappa \cdot x} \kappa \cdot ds \quad (15)$$

and

$$\rho(\kappa) + \rho^*(-\kappa) = \frac{i}{\sqrt{4\pi}} \int_{\kappa \cdot n > 0} e^{-i\kappa \cdot x} \kappa \cdot ds + \frac{i}{\sqrt{4\pi}} \int_{\kappa \cdot n < 0} e^{-i\kappa \cdot x} \kappa \cdot ds \quad (16)$$

$$= \frac{i}{\sqrt{4\pi}} \oint_S e^{-i\kappa \cdot x} \kappa \cdot ds \quad (17)$$

Since the integrand of Eq. (17) is continuous and differentiable on S and in V bounded by S , it follows by Gauss' theorem that

$$\rho(\kappa) + \rho^*(-\kappa) = \frac{i}{\sqrt{4\pi}} \int_V \nabla \cdot (\kappa e^{-i\kappa \cdot x}) dV \quad (18)$$

$$= \frac{\kappa^2}{\sqrt{4\pi}} \int_V e^{-i\kappa \cdot x} dV \quad (19)$$

Introducing the quantity $\Gamma(\kappa)$ defined by

$$\Gamma(\kappa) \equiv \sqrt{4\pi} \frac{\rho(\kappa) + \rho^*(-\kappa)}{\kappa^2} \quad (20)$$

yields for Eq. (19)

$$\Gamma(\kappa) = \int_{\nu} e^{-i\kappa \cdot x} dv \quad (21)$$

Defining the characteristic function $\gamma(x)$ of the scatterer by

$$\gamma(x) \equiv \begin{cases} 1, & x \text{ in } \nu \\ 0, & x \text{ not in } \nu \end{cases} \quad (22)$$

permits the reformulation of Eq. (21) as the three-dimensional Fourier integral

$$\Gamma(\kappa) = \int_{-\infty}^{\infty} e^{-i\kappa \cdot x} \gamma(x) d^3x \quad (23)$$

if the volume ν of the scatterer is finite, then by Eq. (22)

$$\int_{-\infty}^{\infty} |\gamma(x)| d^3x = \nu < \infty \quad (24)$$

it thus follows from Eq. (23) that for finite-size scatterers the three-dimensional inverse Fourier transform of $\Gamma(\kappa)$ exists uniquely,

$$\gamma(x) = \frac{1}{(2\pi)^3} \int_{-\infty}^{\infty} e^{i\kappa \cdot x} \Gamma(\kappa) d^3\kappa \quad (25)$$

which, with the aid of Eq. (20), can be reformulated as

$$\gamma(x) = \frac{1}{4\pi^{3/2}} \left[\int_{-\infty}^{\infty} e^{i\kappa \cdot x} \frac{\rho(\kappa)}{\kappa^2} d^3\kappa + \int_{-\infty}^{\infty} e^{i\kappa \cdot x} \frac{\rho^*(-\kappa)}{\kappa^2} d^3\kappa \right] \quad (26)$$

After replacing κ by $-\kappa$ in the second integral of Eq. (26), we obtain

$$\gamma(x) = \frac{1}{4\pi^{5/2}} \left[\int_{-\infty}^{\infty} e^{i\kappa \cdot x} \frac{\rho(\kappa)}{\kappa^2} d^3\kappa + \int_{-\infty}^{\infty} e^{-i\kappa \cdot x} \frac{\rho^*(\kappa)}{\kappa^2} d^3\kappa \right] \quad (27)$$

$$= \frac{1}{4\pi^{5/2}} \left[\int_{-\infty}^{\infty} e^{i\kappa \cdot x} \frac{\rho(\kappa)}{\kappa^2} d^3\kappa \right] + \left[\frac{1}{4\pi^{5/2}} \int_{-\infty}^{\infty} e^{i\kappa \cdot x} \frac{\rho(\kappa)}{\kappa^2} d^3\kappa \right]^* \quad (28)$$

$$= \frac{1}{2\pi^{5/2}} \operatorname{Re} \int_{-\infty}^{\infty} e^{i\kappa \cdot x} \frac{\rho(\kappa)}{\kappa^2} d^3\kappa \quad (29)$$

Both inverse scattering identities Eqs. (25) and (29) clearly require complete scattering information; namely, knowledge of $\rho(k)$ over all κ space—that is, all frequencies and all aspect angles.

THE FINITE APERTURE INTEGRAL EQUATION

In this section we develop a general, nonsingular, inverse scattering integral equation, solutions to which permit the determination of an appropriate maximum of information about a scatterer from incomplete scattering data. The use of the physical optics approximation is further justified.

In practice $\rho(\kappa)$ is known (measurable) only for an incomplete finite portion of the complete κ space; namely a κ -space aperture consisting of a limited (finite) frequency band and a limited aspect angles band. Furthermore, Eq. (23) is valid only in the physical optics regime (wavelength short compared to the overall size of the scatterer), and hence Eq. (25) must either include fictitious (physical optics scattering data in the Rayleigh regime, which is physically not realizable) low-frequency scattering data $\Gamma(\kappa)$, or no such data at all. We thus turn to the problem of determining what can be deduced about a scatterer $\{\gamma(x)\}$ from such limited high-frequency finite aperture data.

Let $A(\kappa)$ be an *aperture function* defined as

$$A(\kappa) = C(\kappa) W(\kappa) \quad (30)$$

where $C(\kappa)$ is a *characteristic aperture function* defined as

$$C(\kappa) = \begin{cases} 1, & \text{for } \kappa \text{ for which } \Gamma(\kappa) \text{ is known} \\ 0, & \text{for } \kappa \text{ for which } \Gamma(\kappa) \text{ is unknown} \end{cases} \quad (31)$$

and where $W(\kappa)$ is any appropriately chosen (in general nonzero) *aperture weighting function*, subject to the conditions

$$\int_{-\infty}^{\infty} |A(\kappa)| d^3 \kappa < \infty \quad (32)$$

Thus, if the κ -space volume of the aperture is finite, and the aperture weighting function is appropriately chosen, then the three-dimensional inverse Fourier transforms of the aperture and characteristic aperture functions exist uniquely:

$$a(x) = \frac{1}{(2\pi)^3} \int_{-\infty}^{\infty} e^{i\kappa \cdot x} A(\kappa) d^3 \kappa \quad (33)$$

$$c(x) = \frac{1}{(2\pi)^3} \int_{-\infty}^{\infty} e^{i\kappa \cdot x} C(\kappa) d^3 \kappa \quad (34)$$

Thus, by Eq. (25) and the three-dimensional convolution theorem for three-dimensional Fourier transforms, it follows that

$$a(x) * \gamma(x) = \frac{1}{(2\pi)^3} \int_{-\infty}^{\infty} e^{i\kappa \cdot x} \Gamma(\kappa) A(\kappa) d^3 \kappa \quad (35)$$

which by Eqs. (30) and (31) reduces to

$$a(x) * \gamma(x) = \frac{1}{(2\pi)^3} \int_C e^{i\kappa \cdot x} \Gamma(\kappa) W(\kappa) d^3 \kappa \quad (36)$$

where $\Gamma(\kappa)$ is clearly known in the domain of integration C -- the aperture. The right-hand side of Eq. (36) can thus be taken as known, say the known function $g(x)$:

$$g(x) \equiv \frac{1}{(2\pi)^3} \int_C e^{i\kappa \cdot x} \Gamma(\kappa) W(\kappa) d^3 \kappa \quad (37)$$

The three-dimensional inverse scattering problem for a finite aperture thus reduces by Eqs. (36) and (37) to the three-dimensional nonsingular convolution integral equation (a Fredholm integral equation of the first kind)

$$a(x) * \gamma(x) = g(x) \quad (38)$$

This integral equation can be solved numerically by a variety of existing techniques such as the matrix methods of Ritz-Galerkin [Hildebrand, 1965], the associated least-squares best estimate method [Hildebrand, 1965], and the associated moments method of Harrington [1968]; the eigenfunction expression method of Toraldo Di Francia [Wolf, 1970]; leading to so-called "super resolution"; and the κ -space method of Bojarski [1971], which also leads to super resolution. Several closed-form solutions of Eq. (38) for apertures of specific geometry have been obtained by Lewis [1969]; an alternate closed-form solution of Eq. (38) for apertures of general arbitrary geometry is presented in the next section.

A SOLUTION OF THE FINITE APERTURE INTEGRAL EQUATION

Here we develop a closed-form solution to the integral equation of the preceding section. This solution is valid for any arbitrarily shaped finite κ -space aperture (of incomplete frequency and aspect angles scattering data). The full details of a practical frequency and aspect angles band limited right rectangular quasi-conic section aperture are then developed as an example.

The solution of Eq. (38) for $\gamma(x)$ is greatly facilitated by the special properties of $\gamma(x)$ —a priori knowledge that $\gamma(x)$ is a characteristic function of the form Eq. (22)—and the possible judicious choice of the aperture function $F(\kappa)$:

Let the x_3 axis be chosen as passing through the (near) center of the aperture A (fig. 1). Next, let the aperture function $W(\kappa)$ be chosen as

$$W(\kappa) = i\kappa_3 \quad (39)$$

It thus follows from Eqs. (30), (32), (33), and (34), and again the three-dimensional convolution theorem for three-dimensional Fourier transforms, that

$$a(x) = c(x) \otimes \delta(x_1) \delta(x_2) \delta'(x_3) \quad (40)$$

where $c(x)$ is known:

$$c(x) = \frac{1}{(2\pi)^3} \int_C e^{i\kappa \cdot x} d^3\kappa \quad (41)$$

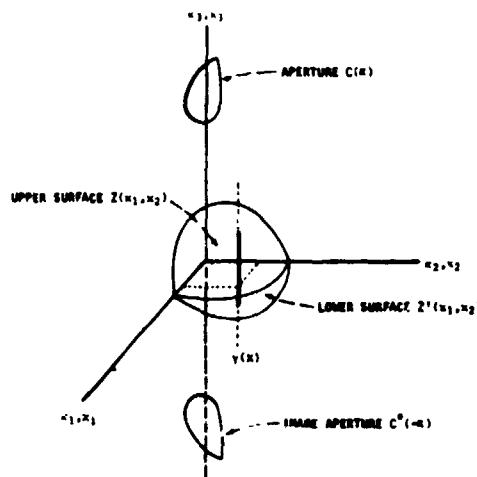


Figure 1.— The aperture and surface coordinate system

It thus follows from Eq. (38) that

$$c(x) \otimes \delta(x_1) \delta(x_2) \delta'(x_3) \otimes \gamma(x) = g(x) \quad (42)$$

which reduces to

$$c(x) \otimes \frac{\partial \gamma}{\partial x_3} = g(x) \quad (43)$$

Since $\gamma(x)$ is a characteristic function of the form (22)—in particular, for fixed values of x_1 and x_2 , $\gamma(x_3)$ is a dual step function in x_3 of unity magnitude and steps at the lower and upper surfaces, say $Z'(x_1, x_2)$ and $Z(x_1, x_2)$, respectively, of the scatterer (fig. 1)—it follows that

$$\frac{\partial \gamma}{\partial x_3} = \delta[x_3 - Z'(x_1, x_2)] - \delta[x_3 - Z(x_1, x_2)] \quad (44)$$

which by Eq. (43) yields

$$c(x) * \delta[x_3 - Z'(x_1, x_2)] - c(x) * \delta[x_3 - Z(x_1, x_2)] = g(x) \quad (45)$$

Examination of Eqs. (20), (26), and (37), and symmetry and physical considerations [the implications of Eqs. (4) and (12)], thus yields for the upper surface $Z(x_1, x_2)$ only

$$c(x) * \delta[x_3 - Z(x_1, x_2)] = -\frac{i}{4\pi^{5/2}} \int_C e^{i\kappa \cdot x} \left[\frac{\kappa_3 \rho(\kappa)}{\kappa_2} \right] d^3 \kappa \quad (46)$$

provided

$$W(\kappa) = W^*(-\kappa) \quad (47)$$

which is assured by Eq. (39).

By the implications of Eqs. (4) and (12), it is physically reasonable that information about the lower surface $Z'(x_1, x_2)$ should only be obtainable from scattering data from the lower image aperture $A^*(-\kappa)$. It is now evident that the introduction of the image aperture served the sole purpose of a mathematical artifice, which permitted the application of Gauss' theorem to Eq. (12), yielding Eqs. (25) and (29); and that knowledge of scattering data in this image aperture is not needed.

The three-dimensional convolution on the left-hand side of Eq. (46), say $\chi(x)$, reduces to

$$\chi(x) \equiv c(x) * \delta[x_3 - Z(x_1, x_2)] \quad (48)$$

$$\equiv \iint_{-\infty}^{\infty} c[x_1 - x'_1, x'_2 - x_2, x_3 - Z(x'_1, x'_2)] dx'_1 dx'_2 \quad (49)$$

Thus

$$\chi(x) = -\frac{i}{4\pi^{5/2}} \int_C e^{i\kappa \cdot x} \left[\frac{\kappa_3 \rho(\kappa)}{\kappa^2} \right] d^3 \kappa \quad (50)$$

where $\chi(x)$ is a resolution density function that is a measure of the location of the upper surface $Z(x_1, x_2)$.

That $\chi(x)$ is indeed such a resolution density function can best be visualized by considering the limiting case of an infinite aperture function $A(\kappa)$ for which $c(x) = \delta(x)$; for such an aperture,

$$\chi(x) = \iiint_{-\infty}^{\infty} \delta(x_1 - x'_1) \delta(x_2 - x'_2) \delta[x_3 - Z(x'_1, x'_2)] dx'_1 dx'_2 \quad (51)$$

$$= \delta(x_3 - Z(x_1, x_2)) \quad (52)$$

by Eq. (49), whereas for a practical realistic aperture of finite κ -space extent, the spatial extent of the nonvanishing portion of $c(x)$, and hence the nonvanishing portion of $\chi(x)$, is still small compared to the size of the scatterer. In fact, this resolution function $\chi(x)$ determines the resolution of the solution Eq. (50), a resolution that can be exceeded only by the super-resolution method mentioned earlier.

A three-dimensional density plot of $\chi(x)$ thus represents the smeared geometrical image of the surface of the scatterer, the spatial extent of the smearing clearly being the spatial extent of $c(x)$ — the resolution.

A best estimate of $Z(x_1, x_2)$ can alternatively be obtained by a variety of correlation [between Eqs. (41) and (49)] methods, employing Fourier transform theory.

THE RESOLUTION

The resolution in x space is clearly the spatial extent of the nonvanishing size of $\chi(x)$; it thus follows from Eqs. (30), (32), and (49) that the resolution in any one direction in x space is the reciprocal of the κ -spatial extent of the (nonvanishing portion of) the aperture $A(\kappa)$ in that one direction.

The finite aperture inverse scattering solution Eq. (50) can clearly be reformulated in a variety of desired practical (radar) spherical coordinate systems. For the particular spherical coordinate system shown in figure 2:

$$\kappa = \kappa \begin{pmatrix} \cos \xi \sin \zeta \\ \sin \xi \\ \cos \xi \cos \zeta \end{pmatrix} \quad (53)$$

$$d^3\kappa = \kappa^2 d\kappa \cos \xi d\xi d\zeta \quad (54)$$

the inverse scattering solution Eq. (50) for the right rectangular, quasiconic section aperture shown becomes

$$\chi(x) = \frac{i}{4\pi^{3/2}} \int_{\kappa_1}^{\kappa_2} \int_{-\xi_0}^{\xi_0} e^{i\kappa x_2 \sin \xi} \int_{-\zeta_0}^{\zeta_0} e^{i\kappa \cos \xi (x_1 \sin \zeta + x_3 \cos \zeta)} \rho(\kappa, \xi, \zeta) \cos \zeta d\zeta \cos^2 \xi d\xi \kappa d\kappa \quad (55)$$

where

$$\kappa_1 = \kappa_0 - \frac{1}{2} \Delta\kappa \quad (56)$$

$$\kappa_2 = \kappa_0 + \frac{1}{2} \Delta\kappa \quad (57)$$

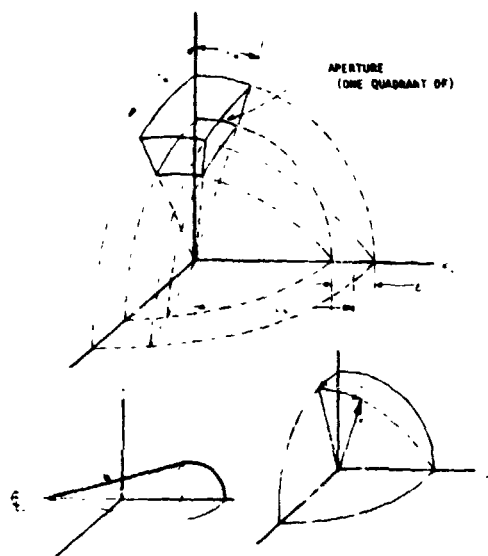


Figure 2. - The mixed coordinate system

where κ_0 and $\Delta\kappa$ are the carrier and bandwidth of the transmitted spectrum respectively:

Examination of Eqs. (41) and (49), after a similar reformulation into this same spherical coordinate system, readily reveals (after small angles approximations) the resolution in range and cross ranges to be

$$\Delta x_1 = \frac{1}{\kappa_0 \Delta \xi} \quad \Delta \xi \equiv 2\xi_0 \quad (58)$$

$$\Delta x_2 = \frac{1}{\kappa_0 \Delta \xi} \quad \Delta \xi \equiv 2\xi_0 \quad (59)$$

$$\Delta x_3 = \frac{1}{\Delta \kappa} \quad (60)$$

This resolution can only be exceeded by the earlier mentioned techniques of super resolution.

Equations (58), (59), and (60) are thus the equations for the parameters of a target resolution radar system.

THE SPECIAL CASES

The special cases of the one- and two-dimensional inverse scattering problem (scattering data restricted to a κ -space line or plane respectively, obtained from a three-dimensional scatterer) can be treated by applying the methods of Lewis [1969] or Bojarski [1967] to equation (30) *et seq.*

Namely, by choosing the characteristic aperture function for the two- and one-dimensional special cases, respectively, as

$$C(\kappa) = C(\kappa_1, \kappa_2) \delta(\kappa_3) \quad (61)$$

$$C(\kappa) = C(\kappa_1) \delta(\kappa_2) \delta(\kappa_3); \quad (62)$$

The three-dimensional convolution integral Eq. (38) reduces to the two- and one-dimensional integral equations

$$a(x_1, x_2) * \beta(r_1, r_2) = g(x_1, x_2) \quad (63)$$

and

$$a(x_1) * \alpha(x_1) = g(x_1) \quad (64)$$

respectively, where β is the *thickness distribution function* in the x_3 direction, α is the *area distribution function* orthogonal to the x_1 direction of the scatterer, and $g(x_1, x_2)$ and $g(x_1)$, respectively, reduce to

$$g(x_1, x_2) = \frac{1}{(2\pi)^2} \int_C e^{i(\kappa_1 x_1 + \kappa_2 x_2)} \Gamma(\kappa_1, \kappa_2) W(\kappa_1, \kappa_2) d\kappa_1 d\kappa_2 \quad (65)$$

$$g(x_1) = \frac{1}{2\pi} \int_C e^{i\kappa_1 x_1} \Gamma(\kappa_1) W(\kappa_1) d\kappa_1 \quad (66)$$

The nonsingular integral equations (63) and (64) can be solved for β and α by any of the previously discussed means.

The further special case of a priori knowledge of the scatterer being a surface of revolution about the x_2 axis (fig. 2) can clearly be treated by the preceding two-dimensional formulation Eqs. (63) and (65) by recognizing that the *profile function* (generatrix of revolution) of the scatterer is $1/2\beta(0, x_2)$, thereby further simplifying Eqs. (63) and (65) after the appropriate modifications. A more direct treatment of the problem of the surface of revolution will be presented subsequently.

The special cases of the aperture $A(\kappa)$ being of certain given geometrical shapes can be treated by applying the method of Lewis [1969] to equation (30) *et seq.*

The special cases of a priori knowledge of the scatterer possessing certain geometrical symmetries can be treated by applying the methods of Bojarski [1967]. We use the case of the scatterer known to be a surface of revolution about the x_2 axis (fig. 2) as an example. For such a scatterer, the monostatic cross section clearly is independent of the longitudinal aspect angle ξ ;

$$\rho(\kappa, \xi, \zeta) = \rho(\kappa, \xi) \quad (67)$$

Furthermore, the profile function $x_3(x_2)$ of such a surface of revolution is given by the function describing the upper surface $Z(x_1, x_2)$ at the plane $x_1 = 0$. Taking the limits of integration over ξ in Eq. (50) as from 0 to 2π , setting $x_1 = 0$ in Eq. (50), and, using the integral representation of the Bessel functions, it follows that

$$\chi(x_1, x_3) = \frac{1}{2\pi^{3/2}} \int_{\kappa_2}^{\kappa_1} \int_{-\xi_0}^{\xi_0} e^{i\kappa x_1} \sin \xi J_1(\kappa x_3 \cos \xi) \rho(\kappa, \xi) \cos^2 \xi d\xi \kappa d\kappa \quad (68)$$

which for small-angle approximations yields the doubly truncated two-dimensional mixed Hankel (Fourier-Bessel)-Fourier transform

$$\chi(x_1, x_3) = \frac{1}{2\pi^{3/2}} \int_{\kappa_1}^{\kappa_2} J_1(\kappa x_3) \int_{-\xi_0}^{\xi_0} e^{i\kappa x_1} \rho(\kappa, \xi) d\xi \kappa d\kappa \quad (69)$$

As expected, only two-dimensional scattering information (in κ and ξ ; i.e., in frequency and one aspect angle, the latitudinal aspect angle) is required for an inverse scattering solution by Eq. (68) or (69).

Solutions are also obtainable for the various combinations of the special cases cited.

CONCLUSIONS

The solutions developed in this paper can be viewed as solutions to the frequency-band-limited and aspect-angles-limited short-pulse synthetic aperture radar imaging (and associated data processing) problem. The solutions presented are based on rigorous electromagnetic scattering and inverse scattering theory applicable to spatially distributed scatterers, yielding real three-dimensional geometrical images; in contrast with conventional two-dimensional radar imaging technique, which is based on the heuristic approach of isolated point scatterers (scattering centers) concepts, yielding so-called "radar images," or maps [Rihaczek, 1969], which most often do not resemble the real geometrical images of the scatterer.

The merits of the solutions presented in this paper over the conventional radar imaging technique deserve the following further examination. The rather unsatisfactory results of the conventional radar imaging technique are essentially the consequence of the heuristic approach to the problem: The approach consists of considering a spatially extended target as a fictitious ensemble of identifiable, stationary, noninteractive, nondispersive, and isotropic point scatterers. From a rigorous electromagnetic scattering point of view, a spatially extended scatterer is not an ensemble of point scatterers, nor are these fictitious point scatterers in principle identifiable, stationary, noninteractive, nondispersive, and/or isotropic. (The point scatterers are not always identifiable by virtue of the so-called registration problem; i.e., that the point scatterers can be continuously, consistently, and correctly identified for various aspect angles. The point scatterers are not always stationary due to the fictitious relocation caused by changing aspect angles.) Furthermore, this technique does not contain basic unique existence considerations, and/or rigorous considerations of the problem of optimizing the results for incomplete aspect angles and/or frequency information availability. The attempts to convert radar images to geometrical images have thus failed for precisely these reasons.

The solutions of this paper alleviate all these objections to the so-called radar imaging technique by the rigorous application of electromagnetic inverse scattering theory, based on rigorous direct scattering theory (and not based on the heuristic model of a spatially extended scatterer as a fictitious ensemble of identifiable, stationary, noninteractive, nondispersive, isotropic point scatterers). It, therefore, avoids the problem of the conversion of radar images to geometrical images, by sidestepping and avoiding the radar image altogether, and addressing itself to the problem of generating actual geometrical images directly from radar data, including unique optimal results from incomplete observation aspect angles and frequency information.

DISCUSSION

Unidentified speaker: If your set C in phase space is bounded in some high-frequency shell then did you say you could get a good determination of the body?

Bojarski: Yes. If the extent of C is such that its transform is small compared to the body size it will map it out for you. The resolution in x space is essentially divided by its size in κ space.

Kay: Part of the limitation imposed on your aperture function is due to the breakdown of physical optics.

Bojarski: The aperture must not be in the Rayleigh or resonance regime.

Unidentified speaker: Do you have a criterion for what limit you should place on the aperture function to guarantee this?

Bojarski: Typically, if the lowest frequency is of the order of either the curvature or 2 or 3 times the Ka number then you get answers within 5 or 6% without noise.

Moses: Have you done numerical modeling to check this?

Bojarski: Yes. We took the Mie series solution for a sphere, truncated the data over a 10% bandwidth, $10ka$ number, $1/10$ rad, to see if we could reconstruct the sphere. We did so within 6% of the profile function.

SOME SPECIAL CASES OF THE ELECTROMAGNETIC INVERSE PROBLEM

V. H. Weston

Purdue University

N73-11616

ABSTRACT

A review of exact techniques for determining the surface of a three-dimensional perfectly conducting body is given, followed by some new results on the uniqueness question concerning the number of measurements that may be required to explicitly determine the surface of the body. It is then shown that the nonhomogeneous but spherically symmetric dielectric electromagnetic case is reducible to a scalar inverse problem that can be treated by known techniques.

INTRODUCTION

A review of exact techniques for determining the surface of a three-dimensional perfectly conducting body is given. This is followed by a discussion on the uniqueness problem for the perfectly conducting case, a subject not well understood but important, since it delineates the necessary set of measurements that may have to be made. With regard to this problem, some new results are presented. Finally the case of a spherically symmetric object is treated- that of a conducting sphere surrounded by a dielectric that is nonhomogeneous in the radial direction. It is shown that this vector inverse problem can be reduced to a scalar inverse problem which can be solved by known techniques.

REVIEW OF THE PERFECTLY CONDUCTING CASE

For a perfectly conducting three-dimensional body, the inverse scattering problem has been treated using continuous wave information. For a fixed direction of incidence, polarization and frequency of the incident wave, the total (scattered plus incident) near field can be computed from knowledge of the far scattered field given over the complete unit sphere of directions. Namely, if the scattered far field has the form in spherical polar coordinates,

$$E^s = \frac{e^{ikR}}{R} E_0(\theta, \phi)$$

the far field data is given by prescribing $E_0(\theta, \phi)$ as a function of the angular variables for $0 \leq \theta \leq \pi$, $0 \leq \phi \leq 2\pi$.

There are various methods of obtaining the near field from the far field data. Wilcox [1956] gives an iterative procedure for determining the near scattered field by expressing it in the form

$$E^s = \frac{e^{ikR}}{R} \sum_{n=0}^{\infty} \frac{E_n}{R^n}(\theta, \phi)$$

and showing that $E_n(\theta, \phi)$ can be calculated through a set of recursive relations involving partial differential operators in the variables (θ, ϕ) , with initial data given by prescribing $E_0(\theta, \phi)$. The expansion is convergent down to the minimum sphere (with center the origin) enclosing the body. However, this domain of convergence can be extended further, under certain conditions as discussed below.

Another approach is to use the plane wave expansion where the near scattered field at a point \mathbf{x} , is given by the integral

$$E^S = \frac{ik}{2\pi} \int_0^{\pi/2 - i\infty} \int_0^{2\pi} e^{i\mathbf{k}' \cdot \mathbf{x}} E_0(\alpha, \beta) \sin \alpha \, d\alpha \, d\beta$$

where $\mathbf{k}' = k(\sin \alpha \cos \beta, \sin \alpha \sin \beta, \cos \alpha)$. Although this integral involves knowledge of the far field data $E_0(\theta, \phi)$ for complex angles of θ , it is shown [Weston *et al.*, 1968], that this can be obtained if $E_0(\theta, \phi)$ is shown only for the real angles for the complete unit sphere of directions. The above expression is convergent in general only for the half plane $z > z_0$, where $z = z_0$ is the tangent plane to the obstacle such that the obstacle lies in the half space $z \leq z_0$. Again for certain cases discussed below, the domain of convergence may be greater, namely a half space $z > z_1$, where $z_1 < z_0$.

Even though the plane representation given above gives the near scattered field in the half-space, by rotation of coordinate system, or by using other forms of the expansion [Twersky, 1962], the near field can be obtained by similar representations for other half spaces; hence in practice, the near field can be obtained in at least the union of all half spaces exterior to the body – outside the minimum convex shape enclosing the body. Although the plane wave representation is mainly useful as a theoretical tool to prove results, it is extremely useful to obtain asymptotic expressions for the near field in the case of high-frequency scattering, even if the far scattered field is not known over the complete unit sphere, but in some sector such as the cap $0 \leq \theta \leq \theta_0$. Stationary phase techniques may be employed to evaluate the integral, and it is shown [Weston and Boerner, 1968] that this is equivalent to tracing the rays (of geometric optics) back to the surface. This asymptotic technique of using the plane wave representation is equivalent in practice to the analog device of holography where the phase and amplitude is measured over a sector usually in the near field.

Another approach which is useful numerically is to expand the near scattered field in terms of the spherical vector wave functions [Stratton, 1941] as follows

$$E^S(\mathbf{x}) = \sum_{n=1}^{\infty} \sum_{m=0}^n a_{eo, mn} m_{eo, mn} + b_{eo, mn} n_{eo, mn}$$

The coefficients a and b (omitting subscripts) are obtained by matching the above expression to the far field data. This involves in practice inverting a matrix to solve a linear system of equations. Some results on the matrices and determinants involved is given by Boerner [1970].

As mentioned above the various expressions for the near scattered field converge in some domain exterior to the body. However, the domain of convergence of the expressions may be extended part way inside the body, if the surface of the body is infinitely differentiable. The domain of convergence is determined by the singularities of the scattered field. As an example, the exact expression for the scattered field by a perfectly conducting sphere centered at the origin (the Mie series) converges and satisfies the free space Maxwell's equations everywhere, except at the origin where it has a singularity. Physically, the scattered field appears to rise from an equivalent source at the origin in contrast to the actual physical currents induced on the sphere.

The location of the domain of convergence is important for two reasons: (1) to know when the scattered field representations may extend part way inside the body, thus providing a means to determine the surface from the total field; and (2) from a practical standpoint, when a finite set of measurements is made, the scattered field has to be approximated by a finite sum, and the knowledge of the domain of the convergence is important for estimating the errors in the scattered field in the vicinity of the surface. This was ignored by Bates [1971] when he assumed that the absolute value of the sum of an infinite series of terms of order ϵ was less than ϵ , where ϵ is a small parameter.

This leads to the question of what conditions must be imposed on the surface for the singularities to lie inside. It is well known that the total field is singular at a discontinuity like an edge. From physical grounds, one must expect that the surface be infinitely differentiable for the singularities to be inside. This was shown for convex portions of the body in Weston *et al.*, [1968], and an algorithm was presented that yields the domain of convergence. The results of the algorithm indicate that for a prolate spheroid, the singularities lie on the axis between the focal points, which agrees with well-known results of R. F. Millar for the elliptic cylinder. Alternative procedures (other than by the exterior representations given above) must be used to obtain the scattered field in cavity portions of the body—that is, portions exterior to the body but inside the domain of convergence of the various exterior representations given above, Weston *et al.*, [1968].

The conducting surface is found by searching for the surfaces for which the tangential components of the total field vanish. The uniqueness question (namely is there more than one such surface for which this is true) is discussed in the next section. From a numerical standpoint, the procedure is to look for the surfaces for which $\mathbf{E} \times \mathbf{E}^* = 0$, where \mathbf{E}^* is the complex conjugate of \mathbf{E} . This is a necessary but not sufficient condition. It implies that the direction of the electric field is real. Another necessary condition is $\mathbf{E} \cdot \mathbf{H} = 0$. Hence, once surfaces are found such that $\mathbf{E} \times \mathbf{E}^* = 0$ and $\mathbf{H} \cdot \mathbf{E} = 0$, then those surfaces for which $\mathbf{n} \times \mathbf{E} = 0$ (\mathbf{n} the unit normal to the surface) can be easily picked out.

UNIQUENESS FOR PERFECTLY CONDUCTING BODIES

We now consider the question of uniqueness for inverse scattering of perfectly conducting bodies. We have pointed out that for a fixed frequency, direction, and polarization of incident wave, the knowledge of the complete scattered far field, led to the determination of the total field, from which the scattering surface is obtained by locating the surface for which tangential \mathbf{E} vanishes. The question of uniqueness arises; namely, is there more than one such surface for which the tangential electric field vanishes. If more than one such surface exists, how many separate measurements must be performed to evaluate the correct surface. By separate measurements, we mean the observation of the complete scattered field for a different incident wave.

We shall restrict ourselves to scattering bodies contained by simply connected closed smooth [Leis, 1967] surfaces. The smoothness condition is not that restrictive physically, since for a finite wavelength, sharp edges exhibit the same characteristics as edges rounded with radii of curvature much less than a wavelength.

Let S_0 be the surface of the scattering body. We know that the scattered field must possess singularities which lie inside S_0 . Hence, if in performing the inverse problem, we find surfaces additional to S_0 such that the tangential electric field vanishes, we can eliminate from consideration, those closed surfaces that do not contain S_0 . For if such a closed surface should lie outside S_0 (but not contain S_0), it will contain no singularities in its interior, and hence cannot represent an obstacle.

There may exist more than one surface S_1, S_2, \dots , on which the tangential electric field vanishes, and which enclose S_0 . At present, however, we need consider only the smallest surface S_1 enclosing S_0 . If the scattering obstacle's surface is infinitely differentiable the singularities of the scattered field lie well inside the surface, and surfaces could exist inside S_0 such that tangential \mathbf{E} vanishes. However, to simplify the analysis, we will consider the case where the additional surface S_1 is outside S_0 . (The other case may be treated in the same way.)

The existence of the vanishing of tangential E on the surfaces S_0 and S_1 implies that the operating frequency is at an eigenfrequency of the volume V enclosed by the surface; i.e., there exist eigenfunctions (E_m, H_m) of Maxwell's equations. In the volume V , the total field is thus a linear combination of these eigenfunctions.

To determine the set of measurements required to eliminate S_1 as a candidate for the surface of the obstacle, the eigenfunctions and their properties will be considered in more detail. The field at any point y in V can be represented in terms of surface currents as follows

$$E(y) = \int \left[i\omega\mu_0 j\phi - j \times \nabla\phi - \frac{i}{\omega\epsilon_0} (\nabla \cdot j) \nabla\phi \right] d\sigma_x \quad (1)$$

where
$$\phi = \frac{e^{ikR}}{4\pi R} \quad R = |x-y| \quad (2)$$

where x is the variable of integration over the surfaces $S = S_0 + S_1$. If it is postulated that the field quantities satisfy the free space Maxwell's equations, in the interior of S_0 and exterior to S_1 , and the radiation condition at infinity, then j and j' are related to the discontinuity in the tangential components of E and H across the surfaces S_0 and S_1 , i.e.,

$$j = -n \times (H^- - H^+)$$

$$j' = n \times (E^- - E^+)$$

where n is the unit outward normal (out of V) to the surfaces, and H^- and H^+ are the respective values on the interior (with respect to V) and exterior of the surfaces.

The eigenfunctions can be represented by setting $j = 0$, and $j' = -2\mu$, giving

$$E(y) = 2 \int_S \mu \times \nabla\phi d\sigma_x \quad y \in V \quad (3)$$

Let y approach the surface S_1 from the interior, then the above equation yields

$$n \times E(y) = \mu + \int_S n(y) \times (\mu \times \nabla\phi) d\sigma_x \quad y \in S$$

and similarly letting y approach the surface S_0 from the interior of V ,

$$n \times E(y) = \mu + \int_S n(y) \times (\mu \times \nabla\phi) d\sigma_x \quad y \in S_0$$

For (E, H) to be an eigenfunction, $n \times E(y) = 0$, hence the necessary and sufficient conditions for $E(y)$ given by Eq. (3) to represent the modes is that μ must satisfy the following integral equations

$$\begin{aligned} \mu + \int_{S_0} n(y) \times (\mu \times \nabla \phi) d\sigma_x &= 0 & y \in S_0 \\ \mu + \int_{S_1} n(y) \times (\mu \times \nabla \phi) d\sigma_x &= 0 & y \in S_1 \end{aligned} \quad (4)$$

The integral operators associated with the above homogeneous integral equations are compact operators [Muller, 1969], with respect to the Banach space of continuous vector-valued functions defined on the smooth surfaces. Because of this, we have from well-known results that the null space is finite — there are only a finite number N of independent eigenfunctions (E_m, H_m) . By introducing the adjoint operator [Muller, 1969] upper-bound estimates on N can be obtained.

We come to the main question: How many additional measurements must be made to eliminate S_1 as being the scattering surface? To try to eliminate S_1 as a candidate, take another set of measurements (different incident wave but same frequency) and find the total field and hence those surfaces for which tangential E vanishes. Again, we will obtain S_0 since it is the obstacle, but we may or may not obtain S_1 . If we don't obtain it, S_1 is eliminated, but if we do happen to obtain S_1 again, we will have to repeat the process, with another set of measurements. Does this repeated process terminate? It will be shown that it does.

Let $H^i(n)$ and $H(n)$ represent the incident and total magnetic fields for the n th set of measurements. It will be shown that if at least $N+1$ separate sets of measurements are made, then it is impossible for tangential $E(n)$ to vanish on S_1 for all $N+1$ sets. For if tangential $E(n)$ vanishes on S_1 for $n = 1, 2, \dots, N+1$, then we have the following integral equation

$$n \times H(n) = 2n \times H^i(n) + 2 \int_{S_1} n \times [(n \times H) \times \nabla \phi]_x d\sigma_x \quad (5)$$

the solution of which exists, and is unique provided that the frequency is not an eigenfrequency of the volume contained by S_1 . Since the tangential component of H is continuous across S_1 , $n \times H$ on S_1 is a linear combination of the N eigenfunctions of the volume contained by S_0 and S_1 . This means that the set $n \times H(n)$, $n = 1, 2, \dots, N+1$, is a linearly dependent set, and from the integral equation (5) that $n \times H^i(n)$, $n = 1, 1, \dots, N+1$, must be a linearly dependent set. This is a contradiction, however, since the incident fields are independent to begin with. Hence, it is impossible for $n \times E$ to vanish on S_1 for the $N+1$ sets of measurements (at the same frequency); thus, at most $N+1$ sets of measurements need to taken to eliminate S_1 . In practice one may need to take only a few separate measurements (at the same frequency).

When more than one surface S_1, S_2, \dots in addition to S_0 exists for which the tangential components of the electric field vanishes, the eigenfunctions associated with the volume enclosed by S_0 and S_2 , where S_2 contains S_1 , must be a subset that belongs to the eigenfunctions of the volume enclosed by S_0 and S_1 since the fields are continuous across S_1 . Hence, the same number $N+1$ of measurements is needed at most to eliminate S_2 as a possible surface of the obstacle.

An alternative approach [Weston and Boerner, 1969], to eliminate the extra surface is to change the frequency. This is based on the result that the eigenfrequencies of a fixed finite volume are a discrete set. Hence, by making a set of measurements at a shifted frequency, S_1 will no longer appear as the result of the new measurement, provided that the frequency has not been shifted to a new eigenfrequency. Theoretically, all one needs is the set of measurements at a frequency and at an infinitesimal increase in the frequency. In numerical practice, however, the problem of eliminating S_1 as a contender for the scattering surface may not be accomplished by such a slight shift in frequency, since the effect of changing frequency may be to shift or perturb the surface S_1 . From a numerical standpoint the shift in frequency must be sufficiently great that the possible shift in the surface S_1 is large enough to detect numerically, taking into account numerical errors.

INVERSE SCATTERING FOR SPHERICALLY SYMMETRIC BODIES

We now consider the case where the body is not only perfectly conducting but also dielectric. To obtain a solution using known techniques for inverse scattering, we consider the case where the body is restricted to be spherically symmetric, in which case the relative dielectric constant ϵ will be taken to be a real function of the radial coordinate R . In particular, the body will be taken to be a perfect conducting sphere of radius R_0 , with dielectric coating of outer radius R_1 .

The spherically symmetric assumption allows one to express the fields in terms of two scalar potentials ψ and χ as follows with time dependence $\exp(-i\omega t)$ assumed but omitted. The spatial dependence of the electric field is given by

$$\mathbf{E} = +i \nabla \times (\psi \mathbf{R}) + \frac{1}{k\epsilon} \nabla \times \nabla \times (\chi \sqrt{\epsilon} \mathbf{R}) \quad (6)$$

which for $R > R_1$ reduces to

$$\mathbf{E} = +i \nabla \times (\psi \mathbf{R}) + \frac{1}{k} \nabla \times \nabla \times (\chi \mathbf{R}) \quad (7)$$

where k is the free space wave number. The assumption of a perfectly conducting surface at $R = R_0$ imposes the following boundary conditions on ψ and χ .

$$\psi(R_0) = 0 \quad (8)$$

$$\left[\frac{\partial(R\sqrt{\epsilon}\chi)}{\partial R} \right]_{R_0} = 0$$

It can be shown that the scalar potentials must satisfy the following equations

$$\nabla^2 \psi + k^2 \epsilon \psi = 0 \quad (9)$$

$$\nabla^2 \chi + \left[-\frac{\rho''}{\rho} + k^2 \epsilon \right] \chi = 0 \quad (10)$$

where

$$\rho' = \frac{d\rho}{dR} \text{ and } \rho = \epsilon^{-1/2} \quad (11)$$

For the inverse scattering problem, we would like to determine how the far field measurements can be separated out to uncouple their dependence on the two potentials, in which case the vector inverse problem is reduced to a scalar inverse problem.

Before investigating this, we need to consider the incident field. Because of spherical symmetry we can fix the direction of the incident wave to lie in the direction of the positive z axis, and polarized in the x direction:

$$E^i = i_x e^{+ikz} \quad (12)$$

It can be shown that this can be split up into two scalar potentials $\psi^i \cos \phi$ and $\psi^i \sin \phi$, where ϕ is the azimuthal angle, as follows,

$$E^i = +i \nabla \times (\psi^i \sin \phi \mathbf{R}) + \frac{1}{k} \nabla \times \nabla \times (\psi^i \cos \phi \mathbf{R}) \quad (13)$$

with

$$\psi^i(\theta, R) = \frac{1}{kR \sin \theta} \left[e^{+ikz} - \frac{1}{2}(1 + \cos \theta)e^{+ikR} - \frac{1}{2}(1 - \cos \theta)e^{-ikR} \right] \quad (14)$$

It then follows that the scalar potentials corresponding to the total field have the form

$$\psi = \psi(R, \theta) \sin \phi$$

$$\chi = \chi(R, \theta) \cos \phi$$

For $R > R_1$, the potentials can be split into scattered and incident parts

$$\begin{aligned} \psi &= \psi^i + \psi^S \\ \chi &= \chi^i + \chi^S \end{aligned} \quad (15)$$

where ψ^i is given by Eq. (14) and $\chi^i = \psi^i$. The scattered components will satisfy the radiation condition and thus will have the form for $R \sim \infty$,

$$\psi^S(R, \theta, \phi) \sim \frac{e^{ikR}}{kR} f(\theta) \sin \phi \quad (16)$$

$$\chi^S(R, \theta, \phi) \sim \frac{e^{ikR}}{kR} g(\theta) \cos \phi \quad (17)$$

From this and Eq. (7), it can be shown that the far scattered field has the form

$$E^S \sim i \frac{e^{ikR}}{kR} \left[E_\theta \cos \phi \hat{i}_\theta - E_\phi \sin \phi \hat{i}_\phi \right] \quad (18)$$

where

$$E_\theta = \frac{f}{\sin \theta} + \frac{\partial g}{\partial \theta} \quad (19)$$

$$E_\phi = \frac{\partial f}{\partial \theta} + \frac{g}{\sin \theta} \quad (20)$$

We now come to the problem of separating out the components f and g from far field measurements. If the far field scattering pattern is measured for $0 \leq \theta \leq \pi$, and a fixed angle ϕ which is not a multiple of $\pi/2$ (otherwise, we would need two fixed values of ϕ), E_θ and E_ϕ can then be obtained as functions of θ . From Eqs. (19) and (20) we obtain the two first-order differential equations

$$E_\theta + E_\phi = \frac{1}{\sin \theta} (f + g) + \frac{d}{d\theta} (f + g)$$

$$E_\theta - E_\phi = -\frac{1}{\sin \theta} (g - f) + \frac{d}{d\theta} (g - f)$$

which on integration yield,

$$\left[(f + g) \frac{(1 - \cos \theta)}{\sin \theta} \right]_{\theta_0}^{\theta} = \int_{\theta_0}^{\theta} \frac{(1 - \cos \theta')}{\sin \theta'} [E_\theta + E_\phi] d\theta' \quad (21)$$

$$\left[(g - f) \frac{(1 - \cos \theta)}{\sin \theta} \right]_{\theta_0}^{\theta} = \int_{\theta_1}^{\theta} \frac{(1 + \cos \theta')}{\sin \theta'} [E_\theta - E_\phi] d\theta' \quad (22)$$

For the fields to be nonsingular on the axis, the following must hold

$$f(\theta) \sim O(\theta), \quad g \sim O(\theta) \quad \text{for } \theta \rightarrow 0$$

and

$$f(\theta) \sim O(\pi - \theta), \quad g \sim O(\pi - \theta) \quad \text{for } \theta \rightarrow \pi$$

hence, the constants of integration to be taken will be

$$\theta_0 = 0 \quad \text{and} \quad \theta_1 = \pi$$

It thus follows

$$f + g = \int_c^\theta \frac{\sin \theta}{\sin \theta'} \frac{1 - \cos \theta'}{1 - \cos \theta} [E_\theta + E_\phi] d\theta' \quad (23)$$

$$f - g = \int_\theta^\pi \frac{\sin \theta}{\sin \theta'} \frac{1 + \cos \theta'}{1 + \cos \theta} [E_\theta - E_\phi] d\theta' \quad (24)$$

The far field components can now be separated out, giving

$$f(\theta) = \frac{1}{2} \int_0^\pi \frac{\sin \theta}{\sin \theta'} \frac{1 \mp \cos \theta'}{1 \mp \cos \theta} [E_\theta \pm E_\phi] d\theta' \quad (25)$$

where the upper and lower sign is taken according as $\theta > \theta'$ or $\theta' > \theta$. Hence once E_θ and E_ϕ are known for $0 \leq \theta \leq \pi$, $f(\theta)$ is determined from Eq. (25). The vector inverse scattering problem now is reducible to the following scalar inverse problem.

Given that

$$\psi(R, \theta, \phi) = \psi(R, \theta) \sin \phi$$

is a solution of the equation

$$\nabla^2 \psi + k^2 \epsilon(R) \psi = 0$$

subject to the boundary condition $\psi = 0$ at $R = R_0$, the problem is to determine $\epsilon(R)$ from the knowledge of the far scattered field $f(\theta)$ where

$$\psi S \sim \frac{e^{ikR}}{kR} f(\theta) \sin \phi$$

for an incident wave ψ^i of the form given by expression Eq. (14).

By expanding the above solution in terms of the associated Legendre function $P_n^1(\cos \theta)$ involving the polar variable, the problem is reduced to a one-dimensional inverse problem involving the radial variable. Hence, with knowledge of the complete scattered field for all frequencies, $\epsilon(R)$ can be computed using the Gelfand and Levitan approach [Newton, 1970].

For the case when $R_0 \rightarrow \infty$, but the thickness of the dielectric layer ($R_1 - R_0$) remains finite, and the measurements of the scattered fields are made in the rear-zone back scattered direction, the three-dimensional inverse problem reduces to the one-dimensional slab problem, which has been treated elsewhere.

This introduces the intermediate case of obtaining a uniform asymptotic expansion to the inverse scattering problem when $R_0^{-1} \rightarrow 0$. If a uniform asymptotic expansion can be obtained, without imposing stringent conditions on $\epsilon(R)$, then correction terms could be obtained to the slab approximation containing the effects of curvature. Present ray-tracing or geometric optics techniques do this, but these are limited to the high-frequency case where $\epsilon(R)$ must be slowly varying.

COMMENTS

We have concentrated on the exact three-dimensional electromagnetic inverse scattering problem. Although in practice the measurements required for the exact inverse problem appear to be out of the realm of practicality, the understanding of the exact problem gives a better insight into the limitations and errors when approximate or asymptotic techniques are employed, and can lead to further refinement of these methods.

DISCUSSION

Kay: Klaus and Muller related the spherical harmonic expansion of the scattered field to the sphere of minimum radius that would enclose the singularities. This sounds close to your results. He did a similar thing for the electromagnetic case.

Weston: Did he relate the behavior where the singularities lie with respect to the body shape?

Kay: What he did was not considered a boundary-value problem as such. He considered the far field that resulted from sources of any kind and related the coefficients of that spherical harmonic expansion to the minimum enclosing sphere. The relationship to your problem has to do with the nature of the induced sources.

Twersky: Bob Miller at NYU, showed for the two-dimensional problem that the corresponding expansions for the circular cylinder the Hankel functions converge down to the axis and for the elliptic cylinder converge down to the focal line.

Unidentified speaker: Have you tried this? There seems to be a stability problem. You have the function only approximately at infinity, how could you determine the body?

Weston: The big problem is to get input data. It has been tried for the spheroid and gave better results than I expected.

ELECTROMAGNETIC SCATTERING

BIBLIOGRAPHY

- Agranovich, Z., and Marchenko, V. A., (trans. by B. D. Seckler), The Inverse Problem of Scattering Theory, Gordon and Breach, New York, 1963.
- Baldwin, D. E., and Kaufman, A. N., Determination of the Density Profile in a Plasma Slab, Phys. Fluids, 12, 1526-1527, 1969.
- Bates, R. H. T., Inverse Scattering for Totally Reflecting Objects, Arch. Rat. Mech., 123-130, 1971.
- Becher, W. D., and Sharpe, C. B., A Synthesis Approach to Magnetotelluric Exploration, Radio Sci. 4, 1089-1094, Nov. 1969.
- Berkowitz, R. S., Modern Radar, 55, John Wiley & Sons, 1965.
- Boerner, W. M., and Vandenberg, F. H., Determination of the Electrical Radius ka of a Spherical Scatterer from the Scattered Field, Can. J. Phys., 49, 1507-1535, 1971.
- Bojarski, N. N., Mie Inverse Scattering for Perfectly Conducting Surfaces of Revolution, Proc. First GISAT Symposium, 1, Pt. 1, p. 39, Bedford, Mass., Dec. 1965, AD 480 500.
- Bojarski, N. N., A Survey of Electromagnetic Inverse Scattering, Syracuse Univ. Res. Corp., Special Projects Lab. Rep., April 1967, AD 813 981.
- Bojarski, N. N., A Differential Equation for the Short Pulse Electromagnetic Inverse Scattering From Surfaces of Revolution, Syracuse Univ. Res. Corp., Special Projects Lab. Rep., Nov. 1968, AD 845 122.
- Bojarski, N. N., The Improvement of Electromagnetic Inverse Scattering Approximations, Syracuse Univ. Res. Corp., Special Projects Lab. Rep., Nov. 1968 AD 845 123.
- Bojarski, N. N., The Determination of the Motion, Orientation, Size and Shape of Axially Symmetric Short Pulse Radar Targets, Syracuse Univ. Res. Corp., Special Projects Lab. Rep., Nov. 1968, AD 845 124.
- Bojarski, N. N., Electromagnetic Short Pulse Inverse Scattering for Discontinuities in an Area Distribution, Syracuse Univ. Res. Corp., Special Projects Lab. Rep., Nov. 1968, AD 845 125.
- Bojarski, N. N., Three Dimensional Electromagnetic Short Pulse Inverse Scattering, Syracuse Univ. Res. Corp., Special Projects Lab. Rep., Nov. 1968, AD 845 126.
- Bojarski, N. N., Three Dimensional Electromagnetic Short Pulse Inverse Scattering with Equatorial Derivative Scattering Data, Syracuse Univ. Res. Corp., Special Projects Lab. Rep., Nov. 1968, AD 845 127.
- Bojarski, N. N., Electromagnetic Inverse Scattering Theory, Syracuse Univ. Res. Corp., Special Projects Lab. Rep., Dec. 1968, AD 711 644.

Bojarski, N. N., K-Space Formulation of the Electromagnetic Scattering Problem, Final Report (Contract F 33615-70-C-1345), Syracuse Univ. Res. Corp., Special Projects Lab. Rep., May 1970, AD 882 040.

Bolinder, F., Fourier Transform in the Theory of Inhomogeneous Transmission Lines, Proc. I.R.E., 38, 1354, 1950.

Bremmer, H., The W.K.B. Approximation as the First Term of a Geometrical Optical Series. Symp. Theory of Electromagnetic Waves, Interscience, New York, 169-179, 1951.

Brown, G. L., The Inverse Reflection Problem for Electric Waves on Non-Uniform Transmission Lines. Ph.D. Thesis, Univ. Wisconsin, 1965.

Burke, J. E., Censor, D., and Twersky, V., Exact Inverse-Separation Series for Multiple Scattering in Two Dimensions, J. Acoust. Soc. Amer., 37, 5-13, 1965.

Butler, J. B., On the Inverse Problem for Ordinary Differential Operators of Even Order, J. Math. App., 26, 142-158, 1969.

Courant, R., and Hilbert, D., Methoden Der Mathematischen Physik, Vol. 1, Julius Springer, Berlin, 1951.

Degasperis, A., Inverse Problem for a Cylindrical Plasma, J. Math. Phys., 11, 3392, 1970.

Dushin, L. A., Pimkin, V. A., Skibenko, A. I., and Fomin, I. P., Determination of the Local Values of Plasma Parameters by the Energy Absorption of Refracting Microwave Beam, transl. by H. Dahlby, Los Alamos Sci. Lab., LA-TR 79-27 July 1969.

Epstein, P. S., Reflection of Waves in an Inhomogeneous Absorbing Medium, Proc. Natl. Acad. Sci. U.S., 16, 627-637, 1930.

Faddeev, L. D., The Inverse Problem in the Quantum Theory of Scattering, J. Math. Phys., 4, 72-104, 1958.

Gelfand, I. M., and Levitan, B. M., On the Determination of a Differential Equation by Its Spectral Function, Dokl. Akad. Nauk. SSSR (N.S.), 77, 557-560, 1951; Izv. Akad. Nauk. SSSR Ser. Mat., 15, 309-360, 1951; Amer. Math. Soc. Transl. Ser. 2, 1, 253-304, 1955.

Gopinath, B., and Sondhi, M. M., Inverse of the Telegraph Equations and Synthesis of Non-Uniform Lines, Proc. IEEE, 59, 383-392, March 1971.

Harrington, R. F., Field Computation by Moment Methods, Macmillan, 1968.

Heim, D. S., and Sharpe, C. B., The Synthesis of Nonuniform Lines of Finite Length-Part I, IEEE Trans. on Circuit Theory, CT-14, 394-403, Dec. 1967.

Hildebrand, F. B., Methods of Applied Mathematics, Prentice-Hall, 1965.

Hylleraas, E. A., Calculation of a Perturbing Central Field of Force From the Elastic Scattering Phase-Shift, Phys. Rev., 74, 48-51, 1948.

- Jacobsson, R., and Martensson, J. O., Evaporated Inhomogeneous Thin Films, Appl. Opt., 5, 29-34, 1966.
- Jost, R., and Kohn, W., Construction of a Potential From a Phase Shift, Phys. Rev., 87, 977-992, 1952.
- Jost, R., and Newton, R. G., The Construction of Potentials from the S-Matrix for Systems of Differential Equations, Il Nuovo Cimento, 1, 590-622, 1955.
- Karp, S. N., Far Field Amplitudes and Inverse Diffraction Theory, in Electromagnetic Waves, R. E. Langer, ed., Univ. Wisconsin Press, Madison, 291-300, 1962.
- Karp, S. N., and Shmoys, J., Calculation of Charge Density Distribution of Multilayers From Transit Time Data, New York Univ. Res. Rep. EM-82, 1955.
- Kay, I., The Inverse Scattering Problem, Res. Rept. EM-74, New York University, Institute of Mathematical Sciences, Division of Electromagnetic Research, 1955.
- Kay, I., On the Determination of the Free Electron Distribution of an Ionized Gas, Research Rep. No. EM-141, New York Univ., Institute of Mathematical Sciences, Division of Electromagnetic Research, 1959.
- Kay, I., The Inverse Scattering Problem When the Reflection Coefficient Is a Rational Function, Commun. Pure Appl. Math., 13, 371-393, 1960.
- Kay, I., Some Remarks Concerning the Bremmer Series, J. Math. Anal. Appl., 3, 40-49, 1961.
- Kay, I., Impedances and Reflection Coefficients for Anisotropic Media, Symp. Electromagnetic Theory and Antennas, Copenhagen, June 25-30, 1962, Pergamon, London, 375-390, 1963.
- Kay, I., Moses, H. E., The Determination of the Scattering Potential From the Spectral Measure Function, Nuovo Cimento, 2, 917-961, 1955; 3, 6-84, 1956; 3, 276-304, 1956.a.
- Kay, I., Moses, H. E., Reflectionless Transmission Through Dielectrics and Scattering Potentials, J. Appl. Phys., 27, 1503-1508, 1956.b.
- Kay, I., Moses, H. E., A Simple Verification of the Gelfand-Levitan Equation for the Three-Dimensional Scattering Problem, Comm. Pure Appl. Math., 14, 435, 1961.a.
- Kay, I., Moses, H. E., V The Gelfand-Levitan Equation for the Three-Dimensional Scattering Problem, Nuovo Cimento, 22, 689-705, 1961.b.
- Keller, J. B., The Inverse Scattering Problem in Geometrical Optics and the Design of Reflectors, IRE Trans. on Antennas and Propagation, AP-7, 146-149, April 1959.
- Keller, J. B., The Inverse Problem of Electromagnetic Scattering by a Metallic Object, Proc. 1st GISAT Symp., Dec. 7-9, 1965.
- Lalor, E., Conditions for the Validity of the Angular Spectrum of Plane Waves, J. Opt. Soc. Amer., 58, 1235-1238, 1968.

- Lalor, E., Inverse Wave Propagator, J. Math. Phys., **9**, 2001-2006, 1968.
- Langer, R. E., An Inverse Problem in Differential Equations, Bull. Amer. Math. Soc., **39**, 814-820, 1933.
- Lavrentiev, M. M., Romenov, V. G., and Vasiliev, V. G., Multidimensional Inverse Problems for Differential Equations, Springer-Verlag Lecture Notes in Mathematics, No. 167, 1970.
- Lax, P. D., Morawetz, C. S., and Phillips, R. S., Exponential Decay of Solutions of the Wave Equation in the Exterior of a Star-Shaped Obstacle, Commun. Pure Appl. Math., **16**, 477-486, 1963.
- Lax, P. D., and Phillips, R. S., The Wave Equation in Exterior Domains, Bull. Amer. Math. Soc., **68**, 47-49, 1962.
- Lax, P. D., and Phillips, R. S., Scattering Theory, Bull. Amer. Math. Soc., **70**, 130-142, 1964.
- Lax, P. D., and Phillips, R. S., Decaying Modes for the Wave Equation in the Exterior of an Obstacle, Commun. Pure Appl. Math., **22**, 737-787, 1969.
- Leis, R., Vorlesungen über Partielle Differentialgleichungen Zweiter Ordnung, Bibliographisches Institut Mannheim, 1967.
- Levinson, N., On the Uniqueness of the Potential in a Schrodinger Equation for a Given Asymptotic Phase, Kgl. Dan. Vidensk Selsk Mat-Fys. Medd., **25**, 1-29, 1949.
- Lewis, R. M., Physical Optics Inverse Defraction, IRE Trans. on Antennas and Propagation **Ap-17**, 308-314, May 1969.
- Marcenko, V. A., Reconstruction of the Potential Energy From the Phase of Scattered Waves, Dokl. Akad. Nauk. SSSR, **104**, 695-698, 1956.
- Marcuvitz, N., The Waveguide Handbook, Gordon and Breach, New York, 1964.
- Mireles, R., The Inverse Problem of Electromagnetic Scattering Theory. I. A Uniqueness Theorem for Cylinders, J. Math. Phys., **45**, 179-187, 1966.
- Moses, H. E., Calculations of the Scattering Potential From Reflection Coefficients, Phys. Rev., **102**, 559-567, 1956.
- Moses, H. E., Solutions of Maxwells Equations in Terms of a Spinor Notation: The Direct and Inverse Problem, Phys. Rev., **113**, 1670-1679, 1959.
- Moses, H. E., and DeRidder, C. M., Properties of Dielectrics From Reflection Coefficients in One Dimension, M.I.T. Lincoln Lab. Tech. Rep. 322, 1963.
- Müller, C., Radiation Patterns and Radiation Fields, Res. Rep. EM-62, New York Univ. Inst. Math. Sci., 1954.

- Müller, C., Electromagnetic Radiation Patterns and Sources, Res. Rep. EM-95, New York Univ. Inst. Math. Sci., 1956.
- Müller, C., Mathematical Theory of Electromagnetic Waves, Springer-Verlag, 1969.
- Newton, R. G., Scattering Theory of Waves and Particles, McGraw-Hill, New York, 1966.
- Newton, R. G., Construction of Potentials From the Phase Shifts at Fixed Energy, J. Math. Phys., 8, 1566-1570, 1967.
- Newton, R. G., Inverse Problem in Physics, SIAM Rev., 12, 346-356, 1970.
- Nirenberg, L., The Weyl and Minkowski Problems in Differential Geometry in the Large, Commun. Pure Appl. Math., 6, 337-399, 1953.
- Nussenzweig, H. M., Phase Problem in Coherence Theory, J. Math. Phys., 8, 561-572, 1967.
- Proc. First GISAT Symp., MITRE Corporation, Dec. 7-9, 1965. (Classified).
- Prosser, R. T., Notes on the Inverse Problem for Radiation Scattering, Proc. Lincoln Labs Symp. 7-12, MIT Tech. Rep. 402, Sept. 27, 1965.
- Prosser, R. T., The Lincoln Calibration Sphere, Proc. IEEE, 53, 1672, Oct. 1965.
- Prosser, R. T., Formal Solutions of Inverse Scattering Problems, J. Math. Phys., 10, 1819-1822, 1969.
- Rihaczek, A. W., Principles of High Resolution Radar, Sect. 9, 331-349, McGraw-Hill, 1969.
- Rydbeck, O. E., A Theoretical Survey of the Possibilities of Determining the Distribution of the Free Electrons in the Upper Atmosphere, Trans. Chalmers Univ. Tech., Gothebur, Sweden, 1, 1942.
- Schelkunoff, S. A., Electromagnetic Theory, Addison-Wesley, New York, 1943.
- Sharpe, C. B., The Synthesis of Infinite Lines, J. Quart. Appl. Math., 21, 105-120, 1963.
- Shmoys, J., Proposed Diagnostic Method for Cylindrical Plasma, J. Appl. Phys., 32, 689-695, 1961.
- Slichter, L. B., Inverse Boundary Value Problem in Electrodynamics, Physics, 4, 411-418, 1932.
- Stratton, J. A., Electromagnetic Theory, Sect. 8.14, 466-467, McGraw-Hill, 1941.
- Titchmarsh, E. C., Eigenfunction Expansions Associated with Second Order Differential Equations, Oxford Univ. Press, London, 1946.
- Twersky, V., Multiple Scattering by Arbitrary Configuration in Three Dimensions, J. Math. Phys., 3, 83-91, 1961.
- Van Bladel, J., Electromagnetic Fields, Sect. 11.8, 360, McGraw-Hill, 1964.

Weston, V. H., and Boerner, W. M., An Inverse Scattering Technique for Electromagnetic Bistatic Scattering, Can. J. Phys., **47**, 1177-1184, 1969.

Weston, V. H., Bowman, J. J., and Ar, E., On the Electromagnetic Inverse Problem, Arch. Ration. Mech. Anal., **31**, 3, 199-213, 1968.

Wheeler, J. A., Scattering and Potential, Phys. Rev., **99**, 630, 1955.

Whittaker, E. G., and Watson, G. N., A Course of Modern Analysis, Cambridge Univ. Press, 1948.

Wilcox, C. H., An Expansion Theorem for Electromagnetic Fields, Commun. Pure Appl. Math., **9**, 115-134, 1956.

Wolf, E., Progress in Optics, **VIII**, Sect. 6, 39-44, North-Holland, 1970.

Wohl, M. R., A Realizability Theory for Smooth Lossless Transmission Lines, IEEE Trans. on Circuit Theory, **CT-13**, 356-363, Dec. 1966.

Youle, D. C., Analysis and Synthesis of Arbitrarily Terminated Lossless Nonuniform Lines, IEEE Trans. on Circuit Theory, **CT-11**, 363-372, 1964.

7. SEISMOLOGY

This chapter is concerned primarily with the earth's interior, although several of the techniques discussed are finding application in space science studies of other solar system bodies such as the Moon. The geophysical inverse problem is concerned with the inversion of seismic data to obtain profiles of parameters which describe the earth's interior. The techniques described here have also found application in radio-occultation studies (Chapter 3), and certain of the techniques rely on the mathematical algorithms developed in the particle scattering (Chapter 5) and electromagnetic scattering (Chapter 6) disciplines.

R. A. Phinney organized and chaired the session devoted to seismology. The paper by Gilbert is a transcription, edited by Dr. Phinney, of the oral presentation.

MONTE CARLO INVERSION OF SEISMIC DATA

Ralph A. Wiggins

University of Toronto

ABSTRACT

N73-11617

The analytic solution to the linear inverse problem provides estimates of the uncertainty of the solution in terms of standard deviations of corrections to a particular solution, resolution of parameter adjustments, and information distribution among the observations. Monte Carlo inversion, when properly executed, can provide all the same kinds of information for nonlinear problems. Proper execution requires a relatively uniform sampling of all possible models. The expense of performing Monte Carlo inversion generally requires strategies to improve the probability of finding passing models. Such strategies can lead to a very strong bias in the distribution of models examined unless great care is taken in their application.

INTRODUCTION

Here we examine the general inverse problem from the point of view of *Monte Carlo inversion* (MCI). First we describe the framework of simple linear inverse problems to establish the nomenclature and purpose of MCI. Next we will see that MCI is a natural alternative to the *analytic inverse* (AI). After a discussion of MCI techniques and difficulties, we present a short discussion of the applicability of MCI for determining the structure of the earth.

This paper is devoted to a review of the principles of MCI and does not include specific descriptions of applications. MCI has been applied to various types of seismic data by several authors. *Keilis-Borok and Yanovskaya* [1967] and *Wiggins* [1969] have inverted body wave observations. *Press and Beihler* [1964], *Press* [1968, 1970a,b], and *V. Keilis-Borok and L. Knopoff* [personal communication, 1968] have inverted various combinations of body wave, surface wave and gravity data. In addition, *Anderssen* [1970] has used MCI for estimating the conductivity structure of the earth.

Let us consider the class of inverse problems dealing with the earth. We assume that we have measured m gross earth data (GED) $g_j, j=1, \dots, m$. *Backus and Gilbert* [1967] define gross earth data as any of the possible observations that may be relevant to a particular solution of the general inverse problem of the earth. We also assume that we can represent the structure of the earth by n discrete parameters $p_i, i=1, \dots, n$. Any inverse problem depends then on exploiting an analytical (computational) relationship between the model parameters and the expected GED

$$g_j = f_j(p_i) \quad j = 1, \dots, m$$

The object is to find set(s) of solution parameters p_i that will minimize the difference between the observed and computed GED $g_j' - g_j$.

If the function $f_j(p_i)$ is linear, we can find a completely analytic description of the class of satisfactory models, the resolution of the parameters and the information distribution among the gross earth data. If the function is not linear, then we are faced with the choice of either linearizing f_j in the neighborhood of a particular solution or resorting to alternate procedures such as MCI.

ANALYTIC INVERSE OF LINEAR PROBLEMS

The treatment of the AI as expressed here is a condensation of a much more extensive survey [Wiggins, 1972]. Other excellent treatments of AI has been stated by Backus and Gilbert [1967, 1968, 1970] and in other papers in this volume.

Let us minimize the difference between g'_j and g_j by expanding the functional relationship in a Taylor's series

$$\Delta g_j = \sum_{i=1}^n \frac{\partial f_j}{\partial p_i} \Delta p_i \quad j=1 \dots m$$

where

$$\Delta g_j = g'_j - g_j$$

$$\Delta p_i = p'_i - p_i$$

We now have a set of m simultaneous equations to solve for the parameters Δp_i . Clearly, one must be able to compute both the function $f_j(p_i)$ and the partial derivatives or *variational parameters* $\partial f_j / \partial p_i$ in order to effect a solution.

Now consider the matrix notation

$$\begin{aligned} \Delta G &= \begin{bmatrix} S_1 \Delta g_1 \\ \vdots \\ S_m \Delta g_m \end{bmatrix} & \Delta P &= \begin{bmatrix} \Delta p_1 \\ \vdots \\ \Delta p_n \end{bmatrix} \\ A_j &= \begin{bmatrix} \frac{\partial f_j}{\partial p_1} & \dots & \frac{\partial f_j}{\partial p_n} \end{bmatrix} \\ A &= \begin{bmatrix} S_1 A_1 \\ \vdots \\ S_m A_m \end{bmatrix} & W &= \begin{bmatrix} w_1 & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & w_m \end{bmatrix} \end{aligned}$$

where the weights $S_i = \sigma / \sigma_{g_i}$ are proportional to the reciprocal of the standard deviations of the GED and the weights w_i are proportional to layer thicknesses or volumes. If we parameterize the model by dividing it into a set of layers or volumes and assigning an average value to each parameter, then this specification of w_i makes the length of the vectors $A_j W^{-1/2}$ invariant to the thicknesses or volumes selected for parameterization.

We can now write the simultaneous equations as

$$m \begin{bmatrix} n \\ A \end{bmatrix} \begin{bmatrix} n \\ \Delta P \end{bmatrix} = \begin{bmatrix} m \\ \Delta G \end{bmatrix}$$

Among these m equations, we will find some k independent relationships between ΔP and ΔG where generally $k \ll n, m$. Note that A has been defined so that indices along rows of A correspond to parameters and indices along columns of A correspond to GED.

Lanczos [1961] has shown that the matrix $AW^{-1/2}$ can be factored as

$$m \begin{bmatrix} n \\ AW^{-1/2} \end{bmatrix} = m \begin{bmatrix} k \\ U \end{bmatrix} \begin{bmatrix} k \\ \Lambda \end{bmatrix} \begin{bmatrix} n \\ V \end{bmatrix}^k$$

where U are column eigenvectors, Λ is a diagonal matrix of eigenvalues, and V are row eigenvectors:

$$AW^{-1/2}V^T = U\Lambda \quad W^{-1/2}A^T U = V^T\Lambda$$

(Note that the V matrix here is the transpose of Lanczos' V matrix.)

The eigenvectors V_i have a simple geometric interpretation. The vector V_1 is that vector whose dot product with the rows of $AW^{-1/2}$ is maximal. That is, $V_1 W^{-1/2} A^T A W^{-1/2} V_1^T$ is a maximum subject to the constraint that $V_1 V_1^T = 1$. Thus V_1 is a geometrical average of the rows of $AW^{-1/2}$. If we remove the component of the A_j parallel to V_1 , then V_2 is a vector that is most nearly parallel to the remainder, and so forth.

We can think of the V_i as a reparameterization of the inverse problem. Each V_i represents a particular linear combination of the old parameters. We define a new parameter vector of length k

$$\Delta P^* = VW^{1/2}\Delta P$$

and, since the V_i are orthonormal

$$\Delta P = W^{-1/2}V^T\Delta P^*$$

Substituting $A = U\Lambda VW^{1/2}$ and $VW^{1/2}\Delta P = \Delta P^*$ into the simultaneous equations $A\Delta P = \Delta G$ we find

$$U\Lambda\Delta P^* = \Delta G$$

which has a least-squares solution

$$\Delta P^* = \Lambda^{-1}U^T\Delta G$$

$$\Delta P = W^{-1/2}V^T\Lambda^{-1}U^T\Delta G$$

Since we have weighted the simultaneous equations by the reciprocal of the GED standard deviations, we can apply the theory of linear least squares estimates [see, e.g., Johnson, 1963, p. 106-110] to find the standard deviations of the parameter adjustments

$$\sigma_i^2 = \text{var}(\Delta p_i^*)$$

$$= \frac{\sigma^2}{\lambda_i^2}$$

$$\sigma_{kj}^2 = \text{var}(\Delta p_j)$$

$$= \sigma^2 \sum_{i=1}^k \frac{v_{ij}^2}{\lambda_i W_j}$$

Generally, there are not k nonzero eigenvalues and $n-k$ zero eigenvalues for the matrix $AW^{-1/2}$. Rather, the eigenvalues approach zero asymptotically. Thus we must choose some k based on the behavior of the standard deviations σ_{kj} . That is, we select an upper bound for the parameter standard deviations and then select k such that the solution standard deviations are not greater than that bound.

Now let us examine briefly what we can say about the solution of the inverse problem based on the above development. We have determined a set of k adjustments to add to the model based on the eigenvectors V_i . The size and standard deviations of the adjustments are determined by the GED and their standard deviations. There are a set of $n-k$ free vectors V_{oi} , $i=1, \dots, n-k$, which are orthogonal to the V_i and whose coefficients are completely undetermined. In the linear problem, we can add any amount of any of the vectors V_{oi} to the model p_i without affecting the quality of the fit at all.

Each of the k adjustments are determined by a linear combination of the gross earth data. The particular combinations are given by the column eigenvectors U_i . That is, U_1 gives the explicit linear combination of gross earth data that determines the coefficient for the eigenvector V_1 . Thus we have also determined a set of k linear relationships between the GED. If any data value falls significantly outside these linear relationships, then our model is improperly specified or the observation is in error.

The eigenvectors V_i and U_i can be further manipulated to estimate the resolving power and information distribution of the problem. First consider the k parameter eigenvectors V_i . The resolving power of a set of GED is determined by finding the smallest sets of adjacent parameters, the average value of which can be determined uniquely. The thicknesses of the corresponding sets of layers are said to be the resolved thicknesses. Clearly, since the k eigenvectors V_i determine the total linear relationship between the parameters, we must explore the resolution properties by considering linear combinations of the V_i . There are a number of ways of finding particular combinations of the V_i that maximize a set of adjacent parameters while minimizing all the others. A consideration of all such combinations will then map the resolution distribution in parameter space. We can make similar manipulations on the eigenvectors U_i to determine adjacent sets of GED that provide equivalent information. All these types of analysis should be performed for a complete description of any particular linear inverse problem.

MONTE CARLO INVERSE

Suppose that we have a particular solution of an AI problem $P_0 = [p_1 \dots p_n]$ together with the eigenvectors V_i and their standard deviations σ_i^* as determined in the last section. Further, let us assign (large) standard deviations σ_{oi}^* to the free vectors V_{oi} . Let us then generate a set of random models P_i , $i = 1, \dots, l$, by adding random amounts of each of the vectors V_i and V_{oi} to the particular solution. The random coefficient for adding each vector will be constrained to have zero mean and a standard deviation of σ_i^* or σ_{oi}^* respectively.

The object of this exercise is to realize that we can recover the eigenvectors V_i and V_{oi} and their standard deviations σ_i^* and σ_{oi}^* from the models. We simply form the model matrix

$$P = \begin{bmatrix} P_1 - P_0 \\ \vdots \\ P_l - P_0 \end{bmatrix}$$

If the standard deviations were distinct values, then the eigenvalues of this matrix will also be distinct and will be proportional to the standard deviations σ_i^* and σ_{oi}^* (arranged by descending numerical size) and the eigenvectors will be equal to the corresponding vectors V_i and V_{oi} . Note that here the best determined vectors have the largest standard deviations as opposed to the analysis of the A matrix in the last section.

If for each of the models P_i we had computed a vector of gross earth data $G_i = [f_1(P_i) \dots f_m(P_i)]$ we could also recover the $U_i F$ vectors from the row eigenvectors of the matrix

$$G = \begin{bmatrix} G_1 - G_0 \\ \vdots \\ G_R - G_0 \end{bmatrix}$$

Hence we see that given an adequate sampling of random models, all of which are solutions of the inverse problem, we can estimate the same quantities that were determined by the analytic inverse: an average solution, the standard deviation of the parameter estimates, the resolving power in parameter space and the information distribution among the GED. In addition, we may be able to establish absolute bounds on the range of parameters.

In principle then, MCI is completely equivalent to AI. Both approaches have basic difficulties. AI requires the ability to compute the partial derivatives $\partial f_j / \partial p_i$ and that these derivatives remain nearly constant throughout the model space. MCI can completely avoid the problem of computing the derivatives. Thus the linearity of the derivatives is of little or no consequence to the solution. We must, however, obtain an adequate sampling of the model space and therein lies the difficulty. Before discussing these difficulties, let us consider a formal statement of the MCI procedure.

In MCI we begin by establishing a space \mathcal{P} of possible models and seek to find the subspace \mathcal{P}_R of allowable models. We will consider here only models which are specified by a set of n parameters p_i . Generally, the bounds of this space are defined by upper and lower limits on the values of the p_i ; however, much more complicated limits can be established. In the simplest implementation of MCI the search consists of the following steps:

1. Generate a random model P_r in the space \mathcal{P} .
2. Compute the corresponding GED G_r .
3. If the differences between the computed and observed GED are less than some fixed multiple of the standard deviations of the GED, then add the model P_r to a list of successful models.
4. Return to step 1.

The direct application of the algorithm just described to most real inverse problems is prohibitively expensive. The ratio of passing models to rejected models is usually very small, and the cost of computing the corresponding GED is often significant. Most researchers attempt to solve this problem by increasing the number of constraints on the model generation procedure or by factoring the problem into sequential segments. If certain GED are affected by only a small segment of the model, then that segment is regenerated until all the GED involved are satisfied. The search then proceeds to a new segment, which again is regenerated until the new GED involved are satisfied, and so forth. This procedure greatly facilitates finding models but may skew the distribution of models radically.

Let us consider a simple example. Our model will consist of eight parameters $p_i, i = 1, \dots, 8$, with the constraints that $p_i = 1, \dots, 8$. Thus the parameter space \mathcal{P} can be represented by the squares in figure 1. The only constraint that we place on the passing models is that the parameters be monotonically increasing: $p_i \geq p_{i-1}$. This restriction can be considered either a further constraint on the model generation or a set of GED. If we take the latter approach, then the MCI consists of generating sets of eight uniformly distributed (pseudo) random integers lying between 1 and 8. We then check to see if these numbers are monotonically increasing. Clearly, the probability of finding such models is very small (approximately 0.003). After much computation, we would find that the probability distribution for the passing models would be as is shown in figure 1(a) [Wiggins, 1969]. Most of the models occur near the line $p_i = i$, and very few occur near the opposite corners. This illustrates the greatest weakness of MCI: the probable distribution of passing models may have little or no relationship to the total range of possible models or to those models which may be physically meaningful. To fully explore the space of possible models for this example, we would need to examine more than 75,000 models!

Of course, we can improve the efficiency of the search by a factor of ~ 300 simply by considering the monotonic criterion as a constraint on the generation of the models rather than as GED. That is, we will only generate models for which $p_i \geq p_{i-1}$. After the generation of each p_i , we will consider only values of p_{i+1} which are $\geq p_i$. If we use a uniform probability distribution for generating each p_i , the distribution of passing models will be as shown in figure 1(b). Nearly all the models lie near the upper limit of the range of possible models. This distribution has even less relationship to the full range of possible models than did the distribution in figure 1(a). The skewing of the model distribution illustrated in this example is a necessary consequence of segmenting the search for passing models. Segmenting can certainly vastly improve the probability of finding passing models, but the distribution of models found may be very different from what would be generated without segmentation. The degree of skewing depends on the number of segments and their interdependence. In this example, each parameter represents a separate segment and the segments are strongly interdependent. Since we know the solution distribution for this example, we know how to alter the probability distribution for generating each p_i so that we can obtain a set of models with the distribution shown in figure 1(a) [Wiggins, 1969]. In general, however, the purpose of MCI is to investigate the distribution of models, and therefore we do not have such preknowledge and cannot compensate for the effects of segmentation.

This discussion can best be summarized as follows. The distribution of models examined is strongly affected by the parameterization and constraints imposed on the generation of the models. A successful MCI procedure must be capable of generating a large proportion of models that span the range of successful models with a relatively uniform probability distribution. Such a desideratum can seldom be realized for models that require a large number of parameters for their specification.

Keius-Borok and Knopoff [personal communication] have proposed an alternative MCI technique called "hedgehog" that avoids some of the problems of model distribution. In their scheme each parameter is allowed to have discrete values only. They perform a random search until a passing model is found. Then they change the search strategy to that of examining all nearest neighbors to passing models until they have explored the entire space of passing models. This approach follows from their belief that one must have all possible passing models available before one can adequately describe the statistics of the inversion.

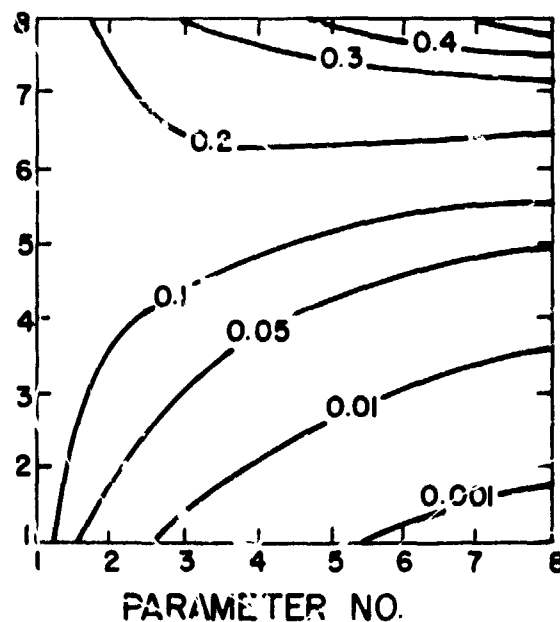
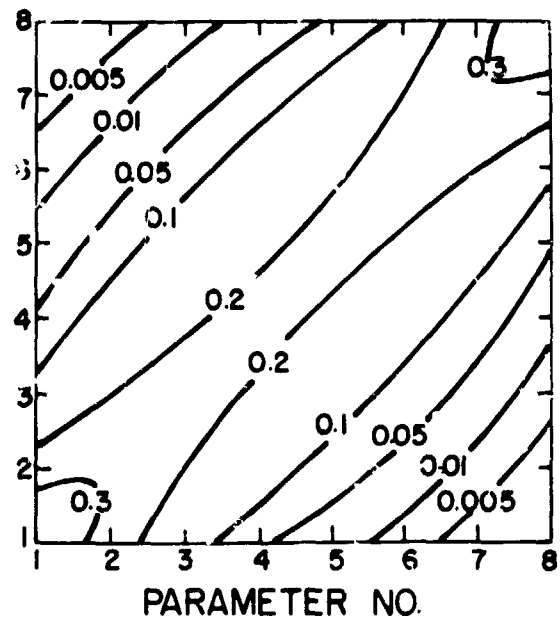


Figure 1.—Probability distributions of random models which consist of eight parameters. Each parameter p_i is restricted to discrete integer values between 1 and 8 such that $p_i \geq p_{i-1}$. The upper portion shows the distribution for a random sampling of all possible models. The lower portion shows a biased distribution produced by improper sampling.

APPLICATIONS

Several years ago, the rule of thumb for inverting GEDs was that the variational parameters for surface waves and free oscillations were sufficiently linear so that AI could be used, and that the variational parameters for body waves were so nonlinear that MCI must be used. This picture is less clear today. The range of models for the upper mantle is sufficiently large that not all these models can be derived by linear perturbations from some "average" model. On the other hand, as shown by Jordan and Johnson and Gilbert later in this chapter, AI can be used for body waves, at least for certain well-behaved parts of the earth's interior. Further, Keilis-Borok and his colleagues have suggested a nonlinear analytic technique that can be extended to place absolute bounds on the range of possible models that can be obtained from body-wave data. Thus, whereas MCI may have become obsolete for inverting body-wave GED, it may be the only practical approach to inverting surface wave GED.

DISCUSSION

Sabatier: Could you give computer time examples for several methods, say, Monte Carlo, hedgehog methods, etc., in seismology?

Wiggins: I can only give a qualitative answer. The hedgehog method is an approach to Monte-Carlo espoused by Keilis-Borok and Knopoff. You only use a random search until you find a particular model that passes and then make the values of the parameters discrete and investigate the space of all possible models that are nearest neighbors to passing models. As regards time for inverting, the hedgehog or straight Monte Carlo are very expensive. Compared to analytically computing the range of possible models it's an impossible comparison. To get the envelope takes a trivial time.

Moses: The use of the amplitudes of the waves suggests, from electromagnetic analogs, that you can go beyond turning points. Is this true?

Wiggins: That's right. The Abel equation approach to inverting has been extended by Gerrard and Markushevich to include low velocity zones. We do not need to deal with models that are monotonically increasing. There is a certain ambiguity in how you construct these zones, however.

Gross: The travel time against distance data is obtained from various stations that are distributed in some fashion and may have entirely different aspects relative to the epicenter. From one earthquake to another you may have waves arriving from different directions at each individual station. Yet you interpret them as if they are all on a line. Is that right?

Wiggins: We make the assumption that the earth is radially symmetric and then subject the observations to a number of incomplete corrections which we think we know.

Gross: How sensitive are the bounds on the model to the radial resolution that you allow?

Wiggins: Qualitatively, the widest bounds are those for models that are cassinated (stairstep). When we force the model to have smoothly varying velocities, we have a reduction of maybe two in the total area of that space. However, it is not clear from physical grounds that the earth has to have a smoothly changing velocity.

BACKUS-GILBERT INVERSION OF TRAVEL TIME DATA

Leonard Evans Johnson*

University of California at San Diego

ABSTRACT

N73-11618

This paper describes the application of the Backus-Gilbert theory for geophysical inverse problems to the seismic body wave travel-time problem. In particular, it is shown how to generate earth models that fit travel-time data to within one standard error and having generated such models how to describe their degree of uniqueness. An example is given to illustrate the process.

INTRODUCTION

In this paper I will describe the application of the Backus-Gilbert theory for geophysical inverse problems [1967, 1968, 1970] to the seismic body wave travel time problem. Given a set of measured travel times for the real earth from a particular source to various surface locations, what is the seismic velocity as a function of depth?

The traditional method of inverting travel time data is due to Herglotz and Wiechert, who observed that in the case of a spherical earth, with the velocity v a function of the radius r alone, if $v(r)/r$ increases with depth, the inversion amounts to solving a form of Abel's integral equation. An elementary solution is given by Jeffreys [1962]:

$$\int_0^{\Delta_1} \cosh^{-1} \frac{p}{p_1} d\Delta = \pi \ln \frac{a}{r_1} \quad (1)$$

where $p = dT/d\Delta$, the derivative of the travel time curve; p_1 is the value of p at $\Delta = \Delta_1$ and is also equal to r_1/v_1 for the seismic ray that has its deepest point of penetration at r_1 ; a = radius of the earth. Now p is a known function of Δ , and Eq. (1) thus determines r_1 corresponding to Δ_1 and hence to r_1/v_1 and so determines v as a function of r .

The Herglotz-Wiechert method is subject to several limitations. First, the method requires that $p = dT/d\Delta$ be known for all values of Δ from 0 to Δ_1 . This requirement cannot be met in the presence of a shadow zone in the interval 0 to Δ_1 in which there are no arrivals of the particular seismic phase being considered. Shadow zones are caused by regions in the earth where the velocity decreases with depth at a rate such that $dv/dr > v/r$. That the earth has such a region, at least for shear waves, was convincingly demonstrated by Dorman, et al. [1960]. Another difficulty with the requirement of a complete travel time curve out to the distance range of interest is that there may be multiplicities in the curve. These multiplicities are caused by regions in the earth where the velocity increases with depth at a rate such that rays penetrating to different depths are refracted to the same location on the surface. This is not a theoretical difficulty provided that all the later arrivals are taken into account. In practice, however, the time intervals between the different arrivals at a given distance in the presence of a multiplicity in the travel time curve are usually very short and cannot be satisfactorily resolved. Second, since the slope of the travel time curve is required, the data are usually smoothed and fitted with a curve, which is then numerically differentiated. This

* Present Address: Cooperative Institute for Research in Environmental Studies, University of Colorado, Boulder, CO.

process strains the data in that a large amount of data closely spaced with respect to epicentral distance is required for accurate slope information. A more fundamental consequence of the finite amount of data available at any given time is that it leads to a gross nonuniqueness in the solution to the problem. *Backus and Gilbert* [1967] prove that the set of earth models which exactly fit the observed data is either empty or an infinite dimensional manifold. Third, the method does not take into account observational errors in the data and the effect that these errors may have on the solution.

The Herglotz-Wiechert method is valuable in the sense that, in the absence of low-velocity zones, it establishes the existence of a solution to the seismic velocity inverse problem. Moreover, *Gerver and Markushevich* [1966, 1967] have proved that if we are given an infinite amount of perfectly accurate travel time data at all distances from sources above and below all the low velocity zones, a solution exists and is unique except in the low velocity zones themselves. It is always comforting to know that a solution to an idealized version of one's problem exists and is unique.

The Backus-Gilbert method of inverting travel time data is not subject to the limitations mentioned above. There is no need of a complete travel time curve, there is no need to fit the data with a smooth curve before inverting, and the Backus-Gilbert method takes observational errors in the data into account. The method exploits the nonuniqueness of the problem to generate solutions that fit the data and, most important, it allows us to make a quantitative statement about the degree of nonuniqueness of the solutions.

INVERSION THEORY

We consider spherically symmetric earth models in which the velocity is only a function of the radius r . We may think of earth models as members of a Hilbert space in which the inner product of a pair of functions \mathbf{m}, \mathbf{m}' is defined as $(\mathbf{m}, \mathbf{m}') = \int_0^1 m(r)m'(r)dr$ and a norm as $\|\mathbf{m}\| = (\mathbf{m}, \mathbf{m}')^{1/2}$. To apply the Backus-Gilbert method we must find, to first order, the change in the travel times δT produced by a small change in the velocity $\delta v(r)$. Following *Backus and Gilbert* [1969], the travel time of a group of nondispersive body waves from a source p_1 to a receiver p_2 is given by

$$T = \int_{p_1}^{p_2} v^{-1} ds \quad (2)$$

where ds is the element of path length. According to Fermat's principle, to first order in δv , we have

$$\delta T = - \int_{p_1}^{p_2} v^{-2} \delta v ds \quad (3)$$

If we introduce spherical polar coordinates, radius r , colatitude θ , and longitude λ ; choose units so that the radius of the earth is at $r = 1$; and consider the ray path from $r = 1, \theta = 0$ to $r = 1, \theta = \Delta, \lambda = 0$, we have [*Bullen*, 1963]

$$\delta T = -2 \int_{r_p}^1 \eta(r)v(r)^{-2} [\eta(r)^2 - p^2]^{-1/2} \delta v(r) dr \quad (4)$$

where r_p is the minimum radius reached by the ray, $\eta(r) = rv(r)^{-1}$, and $p = v^{-1}r \sin i$ is called the ray parameter, which, according to Snell's law, is constant along the ray. The angle i is the angle between the ray and the local radius vector. At $r = r_p$, $i = 90^\circ$ and $p = r_p/v_p$ where $v_p = v(r_p)$.

We can regard T_i , the travel time calculated from a given model $v(r)$ at epicentral distance Δ_i , $i = 1, \dots, N$, as a nonlinear functional of that model. To linearize the problem, we appeal to Fréchet differentiability of the functional T_i . Then, correct to first order in the small quantity $\delta v(r)$, δT_i is a well-defined linear functional and we write Eq. (4) as

$$\delta T_i = \int_0^1 K_i(r) \delta v(r) dr \quad (5)$$

where $K_i(r)$ is the i th data kernel or so-called "Fréchet kernel" and is defined as

$$\begin{aligned} K_i(r) &= -2\eta(r)v(r)^{-2} \left[\eta(r)^2 - p^2 \right]^{-1/2}, \quad r_p^i < r \leq 1 \\ &= 0 \quad 0 \leq r \leq r_p^i \end{aligned} \quad (6)$$

In the nonlinear case, the data kernels are different for different Earth models $v(r)$.

It is often preferable to consider a relative perturbation to the model rather than an absolute one. If we define $G_i(r) = v(r)K_i(r)$ and $m(r) = \delta v(r)/v(r)$, we can write Eq. (5) as

$$\delta T_i = \int_0^1 G_i(r) m(r) dr \quad (7)$$

and $G_i(r)$ is now our data kernel for the model $V(r) = \ln v(r)$, i.e., $m(r) = \delta V(r) = \delta v(r)/v(r)$.

Now, the Backus-Gilbert method requires the data kernels to be square integrable, and it is obvious from Eq. (6) that this is not the case for $G_i(r)$. To remedy this situation, we integrate Eq. (7) by parts once to obtain

$$\delta T_i = m(1)G_i(1) - \int_0^1 G_i(r)m'(r) dr \quad (8)$$

where $G_i(r) = \int_0^r G_i(r) dr$ and $m'(r) = dm(r)/dr$.

Associated with each travel time T_i is an observational error that we assume has zero mean and variance σ_i^2 . If we consider δT_i as the difference between the observed travel time for the real earth $T_i^{obs} \pm \sigma_i$ and the travel time calculated for a particular model T_i^{cal} , and define $\gamma_i \equiv T_i^{obs} - T_i^{cal}$ we can write Eq. (8) as

$$\gamma_i - \sigma_i \leq m(1)G_i(1) - \int_0^1 G_i(r)m'(r) dr \leq \gamma_i + \sigma_i \quad (9)$$

Our objective now is to determine $m(r)$ from Eq. (9) given a finite number of known $\gamma_i \pm \sigma_i$, $i = 1, \dots, N$. Formulated in this way, the problem is ill-posed in the sense that the solution is not unique. We can take advantage of this nonuniqueness if we are willing to accept any $m(r)$ that is a solution to Eq. (9). A method proposed by *Backus and Gilbert* [1967] for choosing a solution is to find the model that satisfies Eq. (9) and is closest in the least-squares sense to a starting model. In particular, we seek to minimize

$$\frac{1}{2} \int_0^1 m'(r)^2 dr + \frac{1}{2} m(1)^2 \quad (10)$$

subject to the constraints Eq. (9). This is a classic problem in the calculus of variations, and if we introduce Lagrange multipliers ν_i and carry out the minimization we have the solution

$$\begin{aligned} m'(r) &= - \sum_j \nu_j G_j(r) \\ m(1) &= \sum_j \nu_j G_j(1) \end{aligned} \quad (11)$$

To determine the ν_i we substitute Eq. (11) into (9) to obtain

$$\gamma_i - \sigma \leq \sum_j \left[\int_0^1 G_i(r) G_j(r) dr + G_i(1) G_j(1) \right] \nu_j \leq \gamma_i + \sigma_i \quad (12)$$

We define the symmetric inner product matrix

$$A_{ij} = \int_0^1 G_i(r) G_j(r) dr + G_i(1) G_j(1) \quad (13)$$

and write Eq. (12) in matrix form as

$$\gamma - \sigma \leq A \cdot \nu \leq \gamma + \sigma \quad (14)$$

We notice that the system of Eq. (12, or (14) would be much simpler to solve if the $G_i(r)$ were orthogonal. Then, the matrix A would be diagonal and Eq. (14) would be a system of linear algebraic equations, one for each ν_i . Since A is symmetric and positive definite there is a linear transformation that diagonalizes A and simultaneously diagonalizes the covariance matrix E_{ij} of the observed data [*Gilbert*, 1971]. Usually, we assume the data have independent standard errors σ_i so that E_{ij} is already diagonal, but this may not be the case. Following *Gilbert* [1971], we let T be the matrix of the diagonalizing transformation and the transformed version of Eq. (14) becomes

$$(T \cdot \gamma - \Lambda^{-1/2}) \leq T \cdot \nu \leq (T \cdot \gamma + \Lambda^{-1/2}) \quad (15)$$

where Λ^{-1} is the transformed diagonal covariance matrix.

When the data kernels are orthogonal, we have from Eqs. (10) and (11)

$$\frac{1}{2} \int_0^1 m'(r)^2 dr + \frac{1}{2} m(1)^2 = \sum_j (T \cdot v)_j^2 \quad (16)$$

Thus, the minimum in Eq. (16) occurs when each $(T \cdot v)_i$ in Eq. (15) is as small as possible. We write the expression $\mathbf{z} = (T \cdot \mathbf{y} - A^{-1/2})$, $\mathbf{u} = (T \cdot \mathbf{y} + A^{-1/2})$, $C = T \cdot v$ and we have

$$0_i \leq c_i \leq u_i \quad i = 1, \dots, N \quad (17)$$

When $0_i > 0$ we take $c_i = 0_i$, when $u_i < 0$ we take $c_i = u_i$ and $c_i = 0$, otherwise. Gilbert [1971] shows that the elements of the transformed covariance matrix E are the inverse eigenvalues of the matrix $\alpha_{ij} = \sigma_i^{-1} A_{ij} \sigma_j^{-1}$. Thus, the transformed data can be ranked according to standard error; that is, large eigenvalues of α correspond to transformed data with small standard error. This allows us to keep track of the growth of the perturbation to the model in Eq. (16) and to reject those transformed data which have standard errors that are larger than we wish to consider.

Once we have found the v_i 's from Eq. (17) we calculate $m'(r)$ from Eq. (11) and integrate once to get $m(r) = \delta v(r)/v(r)$. This integration produces a very smooth perturbation $\delta v(r)$ to the model. We now have a new model $v(r) + \delta v(r)$ from which we can calculate new travel times to compare with the observed travel times. Since we have linearized the problem, the new calculated travel times may not agree with the observed travel times to within one standard error, so we repeat the process outlined above. The process should converge provided the initial values of $\gamma_i = T_i^{obs} - T_i^{cal}$ are not too large.

RESOLVING POWER THEORY

Now that we have a method for generating models that fit the observed data, what can we say about their uniqueness? When we have only a finite amount of data we cannot expect to resolve details of arbitrarily small scale. The best we can hope for is that our data provide us with an estimate of the true value of the model at any particular radius that is in some sense a smoothed or averaged version of the structure around that radius.

Neglecting errors in the data for the moment we have

$$\gamma_i = \int_0^1 G_i(r) m(r) dr \quad (18)$$

If we wish to consider linear averages of $m(r)$, we have

$$\sum_i a_i \gamma_i = \int_0^1 \left[\sum_i a_i G_i(r) \right] m(r) dr \quad (19)$$

We would like to choose the constants a_i in Eq. (19) so that the function in square brackets is localized around some value of the radius, say r_0 . In other words, if we define

$$A(r, r_0) \equiv \sum_i a_i(r_0) G_i(r) \quad (20)$$

we would like $A(r, r_0)$ to resemble a Dirac delta function centered on r_0 ; if this were possible, Eq. (19) would simply give $m(r_0)$, the exact value of $m(r)$ at r_0 . As mentioned above, it is impossible to construct a perfect delta function with only a finite amount of data, but by choosing the a_i carefully we may be able to do a good job of approximating a delta function at r_0 . Then Eq. (19) would give us

$$\sum_i a_i \gamma_i = \int_0^1 A(r, r_0) m(r) dr = \langle m, A \rangle \quad (21)$$

where $\langle m, A \rangle$ is a smoothed or averaged version of $m(r)$ around r_0 . The averaging length would correspond to the width of $A(r, r_0)$ at r_0 .

How can we choose the a_i so that $A(r, r_0)$ is an approximation to a delta function? Backus and Gilbert have investigated several so-called "δ-ness criteria," which are numerical measures of the difference between $A(r, r_0)$ and a delta function. One such measure is

$$d = \int_0^1 \left[\delta(r - r_0) - A(r, r_0) \right]^2 dr \quad (22)$$

with the condition that

$$\int_0^1 A(r, r_0) dr = 1 \quad (23)$$

The a_i are found by minimizing Eq. (22) subject to the condition (23)

The criterion in Eq. (22) is not the most appropriate one to use in the travel time problem because the data kernels $G_i(r)$ in the expression (20) are not square integrable. However, the

$$G_i(r) = \int_0^r G_i(r) dr$$

are square integrable and so an obvious extension of Eq. (22) is

$$\delta = 12 \int_0^1 \left[H(r - r_0) - \int_0^r A(r, r_0) dr \right]^2 dr \quad (24)$$

where $H(r - r_0)$ is the unit step function. Suppose $A(r, r_0)$ were the "box car" function—that is,

$$\begin{aligned} A(r, r_0) &= 1/\ell & |r - r_0| < \ell/2 \\ &= 0 & |r - r_0| > \ell/2 \end{aligned}$$

then $\Delta = 2$. Thus, we see that Δ is a measure of the width of our averaging function $A(r, r_0)$, which we will call the "spread." The criterion Eq. (24) is an example of a linear "quelling" of a discontinuous linear functional described by *Backus* [1970]. Once we have found the constants a_i from the minimization of Eq. (24) subject to the condition Eq. (23) we can calculate the smoothed or averaged version of $m(r)$ from Eq. (21). If $m(r)$ has any fine-scale structural detail with wavelength smaller than our measure Δ of the spread of the averaging function $A(r, r_0)$, we will not be able to resolve it. This is an inevitable consequence of the finite amount of data we have available and the spread Δ is a quantitative measure of the uniqueness of our inversion.

So far, in discussing resolving power, we have assumed perfectly accurate data. What effect do the observational errors in the data have on the resolving power? If the data have observational errors ΔT_i , then it is easy to see from Eq. (20) that these errors produce an error $\Delta \langle m, A \rangle$ in our averaged value of $m(r)$ at r_0 given by

$$\Delta \langle m, A \rangle = \sum_i a_i \Delta T_i \quad (25)$$

It is clear that a highly localized average of $m(r)$ at r_0 is not very useful if the associated error in the average in Eq. (25) is large. We would be willing to accept an average with a slightly larger spread if we could reduce the error in Eq. (25) appreciably.

Now we don't know the ΔT_i exactly, but we assume we know something about their statistics. In particular, we assume they have zero mean and finite variance: that is, we assume that the covariance matrix

$$E_{ij} = E \{ \Delta T_i \Delta T_j \}$$

exists where $E \{ \cdot \}$ means expected value. We take the square root of the variance of $\Delta \langle m, A \rangle$ as an estimate of the error we make in calculating $\langle m, A \rangle$ with erroneous data, call it e . We have

$$e^2 = \sum_i \sum_j a_i a_j E_{ij} \quad (26)$$

or in linear operator notation

$$e^2 = \mathbf{a} \cdot \mathbf{E} \cdot \mathbf{a} \quad (27)$$

For a precise definition of the notation see *Backus and Gilbert* [1970].

Similarly, after some algebraic manipulation, we can write Eq. (24) as

$$\Delta = \mathbf{a} \cdot \mathbf{S} \cdot \mathbf{a} - 2\Gamma \cdot \mathbf{a} + k \quad (28)$$

where

$$\begin{aligned} S_{ij} &= \int_0^1 G_i(r) G_j(r) dr \\ \Gamma_j &= \int_0^1 H(r-r_0) G_j(r) dr \\ k &= 1 - r_0 \end{aligned} \quad (29)$$

and the condition (23) as

$$\mathbf{a} \cdot \mathbf{G} = 1 \quad (30)$$

Our object is now to find the \mathbf{a} that minimize both the spread given by Eq. (28) and the error in the average of $m(r)$ given by Eq. (27) while satisfying the condition (30). Clearly, we cannot minimize Eqs. (28) and (27) with the same set of \mathbf{a} but *Backus and Gilbert* [1970] show that we can minimize a linear combination of the two. Consider the combination

$$q = \delta \cos \theta + \epsilon^2 \sin \theta \quad (31)$$

where θ is a parameter that runs from 0 to $\pi/2$. When $\theta = 0$ we are minimizing spread and when $\theta = \pi/2$ we are minimizing the error in the average of $m(r)$. *Backus and Gilbert* [1970] prove that as θ goes from 0 to $\pi/2$ the curve $\epsilon^2(\delta)$, called the tradeoff curve of error versus spread, is a monotonically decreasing function of δ ; that is, we can lower the error in the estimate of $\langle m, A \rangle$ by willingness to accept a larger spread.

Minimizing Eq. (31) leads to a system of equations for the \mathbf{a} of the form

$$\mathbf{Q} \cdot \mathbf{a} = 12\Gamma \cos \theta \quad (32)$$

where

$$Q_{ij} = 12S_{ij} \cos \theta + E_{ij} \sin \theta \quad (33)$$

There will be a different set of \mathbf{a} for each value of θ on the tradeoff curve. Once again, the system Eq. (32) would be much simpler to solve if \mathbf{Q} were a diagonal matrix; and once again we can use the diagonalizing transformation developed by *Gilbert* [1971] to simultaneously diagonalize \mathbf{S} and \mathbf{E} .

To summarize briefly, at each radius r_0 where we wish to calculate $\langle m, A \rangle$ we minimize Eq. (31) subject to the condition (30) to determine the \mathbf{a} 's at each point on the tradeoff curve. At each point on the tradeoff curve at r_0 , we calculate the spread δ by Eq. (28) and the error in $\langle m, A \rangle$ at r_0 by Eq. (27).

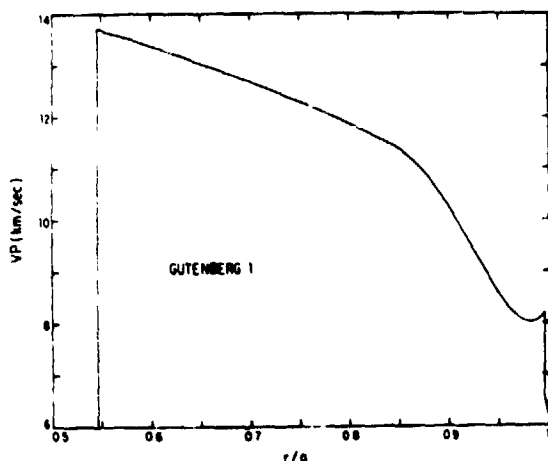
RESULTS

Most of the results of the inversion and resolving power calculations using body wave travel time data will be published elsewhere. However, one example is given to illustrate the method.

Figure 1 shows a starting model for the inversion of a set of data consisting of 70 P-wave travel times from $25^\circ - (1^\circ) - 94^\circ$ taken from *Evernden and Clark* [1970], and 30 P_cP travel times from $15^\circ - (2^\circ) - 73^\circ$ taken from *Herrin et al.*, [1968].

Figure 2 shows the result of the fifth iterate in the inversion process that fits the data to within one standard error (standard errors of 0.5 sec were taken for the P and P_cP data). *Evernden and Clark's* travel time curve (EC-SI-1 in their paper) consists of a series of matched straight line segments. As a result, there is considerable detail in the final model which fits this data. The perturbations in velocity evident in figure 2 between $r/a = .55$ and $r/a = .85$ are on the order of 2% over distances of 100-200 km or greater. The question now is, are these details resolvable?

Note that we have not made a statement about the real earth yet. *Backus and Gilbert* [1968, 1970] show that only if the real earth is linearly close to our model, in the sense in which we have linearized the problem, will it have the same linear averages as our model.



EC-SL-1 (P)
M (PeP)

r/a	VP (km/sec)
0.5	13.8
0.6	13.5
0.7	12.8
0.8	12.0
0.9	10.5
0.95	8.2
1.0	8.5

Figure 4 is a picture of the step functions constructed from linear combinations of data kernels using the criterion in Eq. (24). The step functions are pictured at various points from $\theta = 0$ to $\theta = \pi/2$ along the tradeoff curve at $r/a = 0.7$ for the model shown in figure 2. Recall that near $\theta = 0$ we are minimizing spread so the step functions are very steep and the resulting "delta functions" would be very narrow. Near $\theta = \pi/2$ we are minimizing error so that the step functions are more gradual and the resulting "delta functions" are wider; that is, they have more spread.

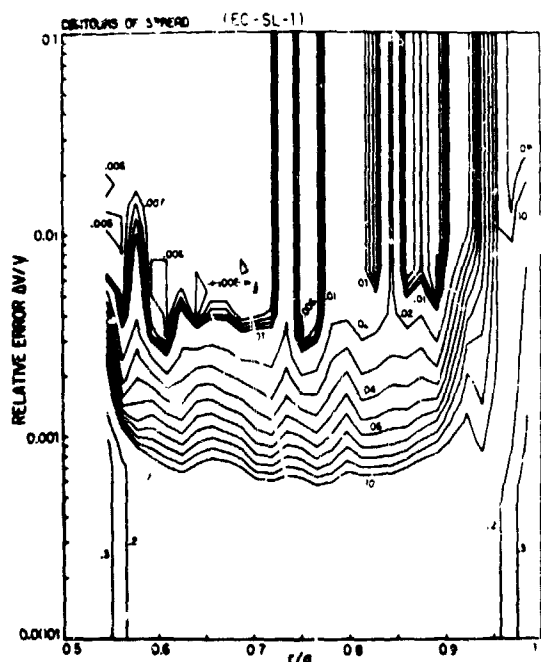


Figure 3. — A contour map of the spread of the optimal averaging function $A(r, r_0)$ as a function of r_0 and the relative error in $\langle m, A \rangle$. The contours of spread are given in fractions of an earth radius.

ACKNOWLEDGMENTS

I am grateful to Professors Freeman Gilbert and Robert Parker for their guidance and encouragement during the course of this work.

Part of this research was supported by the National Science Foundation under contract No. NSF-GA-1225.

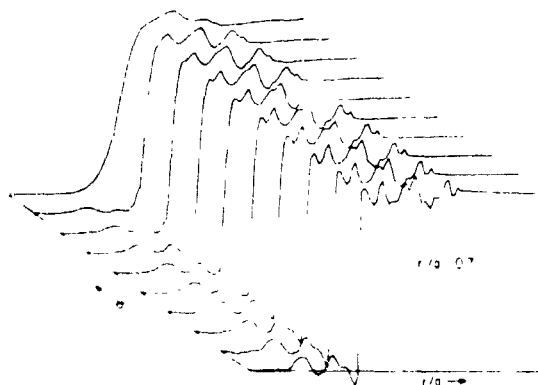


Figure 4. Step-functions constructed from linear combinations of data kernels using the criterion in Eq. (24). The step-functions are pictured at various points from $\theta = 0$ to $\theta = \pi/2$ along the tradeoff curve at $r_0/a = 0.7$ for the model shown in figure 2.

DISCUSSION

Parker: The method you are using, because you've linearized and made various approximations, doesn't depend strictly on the data available. You can map the data in any way you like that is more convenient, as you've done. You can map it into a more convenient form to avoid triplications; for example, you can exploit the fact that you know that the curve of the matched data is piecewise continuous. Thus, you incorporate assumptions and remove the difficulties of interpretation. Having a general method available means you aren't dependent on the type of data available. You can map it, always under the assumption that linearity is a valid assumption.

THE BACKUS-GILBERT METHOD AND ITS APPLICATION TO THE ELECTRICAL CONDUCTIVITY PROBLEM

Robert L. Parker

University of California at San Diego

ABSTRACT

N73-11619

The theory of Backus and Gilbert gives a technique for solving the general linear inverse problem. Observational error and lack of data are shown to reduce the reliability of the solution in different ways: the former introduces statistical uncertainties in the model, while the latter smooths out the detail. Precision can be improved by sacrificing resolving power, and vice versa, so that some compromise may be made between the two in choosing the best model. Nonlinear inverse problems can be brought into the domain of the theory by linearizing about a typical solution.

The inverse problem of electrical conductivity in the mantle is used to illustrate the Backus-Gilbert technique; an example of the tradeoff diagram is given.

INTRODUCTION

The most general theory for handling inverse problems that we have today is that of Backus and Gilbert [Backus and Gilbert 1967, 1968, 1970; Backus 1970a,b,c]. This work has gained the reputation of being mathematically very difficult. However, it is not the basic concepts that are difficult. Rather, to make their exposition rigorous and general, the original authors needed powerful mathematical weapons that are not in the armory of the average geophysicist. The situation is analogous to that of quantum mechanics: Von Neumann rigorized the subject by using the theory of unbounded operators and spectral theory in Hilbert space, but undergraduates grasp the basic ideas of quantum mechanics without this apparatus. Backus and Gilbert also use Hilbert space to develop their theory, but we shall not need any such profound mathematics here.

What is an inverse problem? Suppose we have made measurements of the magnetic field of the earth. What does this tell us about the magnetization of surface rocks? Or given a set of normal mode frequencies (the frequencies of free oscillation of the earth observed after the largest earthquakes), we would like to find the density and seismic velocities as a function of radius. These are typical geophysical inverse problems; they are attempts to infer properties of the interior of the earth, which is inaccessible to us, from measurements made at the surface. The general procedure in such problems is to construct a model of the earth that possesses the same external observables as those measured in practice. With modern computers this is no longer a difficult task. The real difficulty is in knowing how well the data determine the property we are seeking. For example, it is well known that measurements of the external gravity field of the earth do not determine uniquely the density within it: there are infinitely many structures, widely differing from each other, which all give rise to the same field. This is the sort of problem that Backus and Gilbert have solved and, what is more, to which they have brought a great deal of insight.

THE BACKUS-GILBERT TECHNIQUE

The Data

We first define the observations to be the N real numbers γ_j , $j = 1, 2, \dots, N$, where N is finite. Clearly, in the normal mode example, we will always have a limited number of measurements. But what about the case of

magnetic data recorded on a strip chart? There we have a continuous function. How can we pretend that an adequate representation is obtained by N numbers? There are two ways to answer this objection. The first is the frivolous suggestion that since we now always process our observations on a computer, which cannot hold an infinite number of samples, we must be satisfied with something less. This suggestion dodges the issue because we have appealed only to the practical nature of data processing, one might ask whether continuous records should be digitized so closely that the number of samples is as large as the available computer memory permits.

It is more pertinent to appeal to the practical nature of measurement. All recording instruments have a limited bandwidth, and all records are finite in length. It is clear that our continuous chart recording could be Fourier analyzed (there is no problem about existence of the Fourier series here), and coefficients corresponding to frequencies higher than the instrument cutoff must be useless. Thus, a finite number of numbers (the N accepted Fourier coefficients) represent the original continuous curve adequately and can be used as the data γ_j . In time-series analysis, this is called choosing the proper Nyquist frequency to sample all the information.

DELTA FUNCTIONS

Suppose we now have our data, and for the present assume they are exact—no experimental error. Let us take the problem of determining a property in the earth as a function of radius, assuming no angular dependence. Furthermore, we initially consider a *linear* inverse problem; this means the observable depends in a linear way on the property. Under these assumptions we can write

$$\gamma_j = \int_0^a m(r) G_j(r) dr \quad j = 1, 2, \dots, N$$

where $m(r)$ is the property and $G_j(r)$ is called a data kernel, one for each observable. A simple but instructive example is to make $m(r)$ the density and γ_j the radial component of gravity at radius R_j from the center of the earth ($R_j \geq a$). Then

$$\gamma_j = \int_0^a \frac{4\pi G r^2}{R_j^2} \rho(r) dr$$

where G is Newton's gravitational constant. We see at once

$$G_j = \frac{4\pi G r^2}{R_j^2}$$

and the problem is, of course, linear.

Now let us return to the more general radial property $m(r)$. The only thing we know about m is the set of measurements γ_j . How can we localize this information to points within the earth? Backus and Gilbert make the following suggestion: consider a linear combination of γ_j given by

$$L = \sum_{j=1}^N a_j \gamma_j$$

From our linearity assumption we have

$$L = \int_0^a \left[\sum_{j=1}^N a_j G_j(r) \right] m(r) dr$$

Now suppose we could choose the constants a_j so that the function in square brackets was a Dirac delta function centered on r_0 . Then L would simply be $m(r_0)$, the property we want at the position r_0 . In general, it is impossible to get ideal delta functions, but by choosing our coefficients a_j carefully we might be able to find a function that is concentrated strongly at r_0 . How do we choose a_j to provide approximations to a delta function at various radii? Backus and Gilbert define a numerical measure of "difference from a delta function" and minimize this number by varying the coefficients a_j . For example, they investigated these two measures:

$$D_1 = \int_0^a [\delta(r - r_0) - F(r)]^2 dr$$

$$D_2 = \int_0^a F(r)^2 (r - r_0)^2 dr$$

with $\int_0^a F(r) dr = 1$. The measure D_1 is essentially a mean square deviation, the other is a more sophisticated criterion, but I think one can see that if D_1 or D_2 can be made small, the function $F(r)$ will have a large peak at r_0 and unit area under it. Other criteria are possible but the two above have proved useful because they yield simple equations for the coefficients a_j .

Having decided on a criterion and performed the minimization we will obtain for every radius r_0 a set of a_j and corresponding to it an estimate of the property $m(r_0)$ given by

$$\langle m(r_0) \rangle = \sum_j^N a_j \gamma_j$$

So the information we gain about the property m is a smoothed version of the actual structure; the real profile has been convolved with the best "delta function" we can find from linear combination of the G_j . If there is detailed structure that has wave lengths less than the width of our "delta function" at a particular depth, our measurements cannot reveal it. In other words the finite number of data (perfectly accurate though they are) give us a blurred picture of the earth. We have thus discovered the extent to which our inversion is unique.

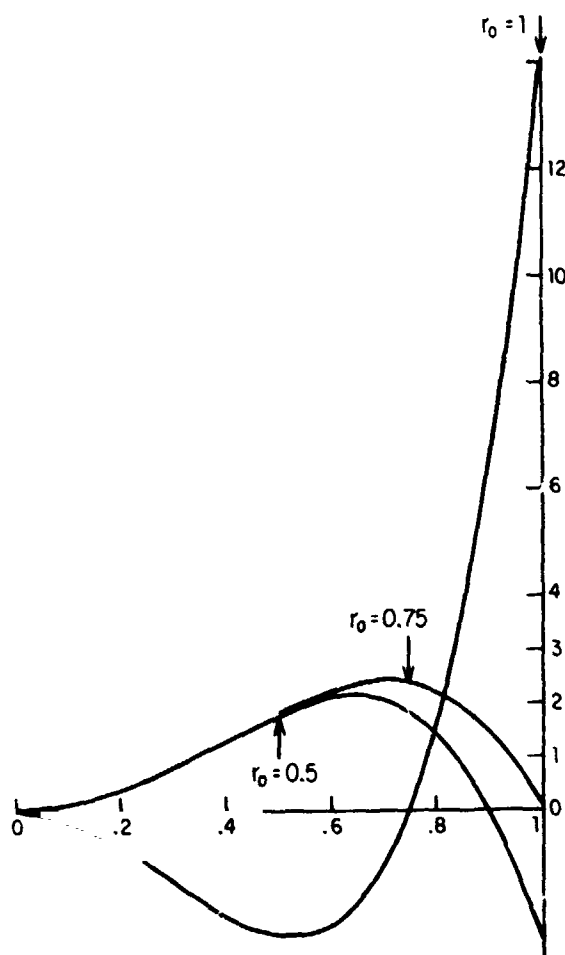


Figure 1.— Three delta functions for density using gravity and moment of inertia.

Returning now to the example of gravity, we can see that measurements at different R_j yield no new information about $\rho(r)$, because any linear combination of the G_j is still proportional to r^2 . This is the best (and only!) approximation to a delta function that we can make. Thus, rather trivially, we have confirmed the statement made earlier: any radially structured model with the correct mass satisfies the data. Suppose now we include the moment of inertia as a datum γ_2 , while γ_1 is a single gravity measurement:

$$\gamma_2 = \int_0^a \frac{8\pi}{3} r^4 \rho(r) dr$$

so this datum is linear in ρ . Without going into details we can see that linear combination of the two data kernels will allow us some sort of approximation to a peaked function. Using the D_1 criterion I calculated some curves (see fig. 1). We are not getting very good resolution with these data. If experiments were available that gave data whose kernels were proportional to r^6 , r^8 , r^{10} , etc., it can be proved that arbitrarily good delta functions can be constructed from them by means of the Weierstrass approximation theorem [Courant and Hilbert, 1966]. Unfortunately these data do not exist!

The Role of Experimental Error

So far all our data have been precise: we have seen that with only a finite number of measurements our knowledge must be imperfect even in this ideal situation. In practice, each of the measurements will be subject to error, whose rough size the experimenter usually estimates.

The property averages $\langle m(r_0) \rangle$ are simply linear combinations of the data γ_j so that it is easy to find the associated errors of their values. Now we have two measures of imprecision both of which we should like to make as small as possible—the error of the estimate and its resolution (defined to be the half width of the delta function for example). Can the a_j be chosen to minimize both at once? In fact this is impossible: some compromise must be accepted. Let us use the D_2 criterion for delta function quality; it is possible to define the resolution width s to be equal to $12D_2$ (this has dimensions of length). After some manipulation one can show that

$$s = \sum_i \sum_j a_i a_j S_{ij}$$

where S_{ij} is a matrix that depends on the G_j and r_0 . Similarly, the error estimate ϵ of $\langle m(r_0) \rangle$ is given by a quadratic form

$$\epsilon^2 = \sum_j a_j^2 \sigma_j^2$$

where r_j has standard error σ_j (it is assumed here that the errors are statistically independent). Since we cannot minimize ϵ^2 and s with one set of a_j , let us agree to minimize instead a composite quadratic form

$$Q(a_j) = (1 - w)s + w\epsilon^2$$

where w is a weighting factor that, when 0, gives all the weight to resolution (and, of course, yields the solution we discussed earlier), and when 1, gives all the weight to error. *Backus and Gilbert* [1970] show that as w varies from 0 to 1, and we minimize Q by choosing the appropriate set of a_j , the resolution length increases and the error decreases. In other words, one can improve the error estimate of a property average by sacrificing its resolving power and vice versa. The "best" choice of w is impossible to give: it depends on what one is trying to say about the earth. Therefore, to solve a real problem completely one must give the relationship between error and resolution at every radius.

The situation outlined above is analogous to that described by the quantum mechanical uncertainty principle; it is impossible to give a perfectly accurate property value at a perfectly defined position. We can improve one at the expense of the other, but in the case of linear inverse theory neither the accuracy nor the resolution can be refined indefinitely: the inadequacy of our data puts bounds on both. We shall look at an example of a "tradeoff" diagram in discussing the electrical conductivity problem.

Nonlinear Extension

Most (but not all) interesting geophysical inverse problems are nonlinear. The observable data are nonlinear functionals of the earth property in question, and consequently there is usually no explicit relation between γ_j and $m(r)$. Nonlinear analysis is notoriously difficult, and the solution proposed by Backus and Gilbert is simply to linearize the problem. A model of the earth is found that fits the data well (how this is done will not concern us): call it $m^{(0)}(r)$ and the observables it predicts $\gamma_j^{(0)}$. Then it can be shown, by a process rather like Taylor series expansion, that it is usually possible to write

$$\gamma_j - \gamma_j^{(0)} = \int_0^a [m(r) - m^{(0)}(r)] G_j(r) dr + 0 \int_0^a [m(r) - m^{(0)}(r)]^2 dr$$

If the second term is neglected, this is a new linear problem with $\gamma_j - \gamma_j^{(0)}$ as the N observables and $m(r) - m^{(0)}(r)$ as the required property.

Such linearization is clearly an approximation, and even if it is a valid one, we have no guarantee that a "base" model other than $m^{(0)}(r)$ does not exist that is outside the scope of the linear description. Nonetheless, a linear description of the uncertainties in inverse models gives us a much clearer understanding than we had before.

THE ELECTRICAL CONDUCTIVITY PROBLEM

Physical Background

What are the surface influences of electrically conducting material at great depths within the earth? Geomagnetic variations are modified by the presence of a conductor in the following way: magnetic field fluctuations cause eddy currents to flow in the conductor, and these currents themselves give rise to magnetic fields. Remarkably, potential theory gives a method whereby a surface observer can distinguish between the driving field, caused outside the earth, and the induced field appearing in response to the first. The method requires the surface field to be

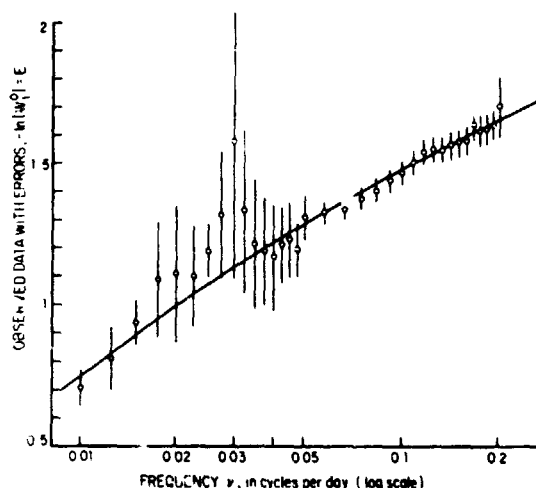


Figure 2. — Thirty-five data points taken from Banks [1969] showing the amplitude ratio of the vertical to horizontal field in the frequency range 0.2 to 0.01 cycle per day. Solid line result of theoretical iterative model.

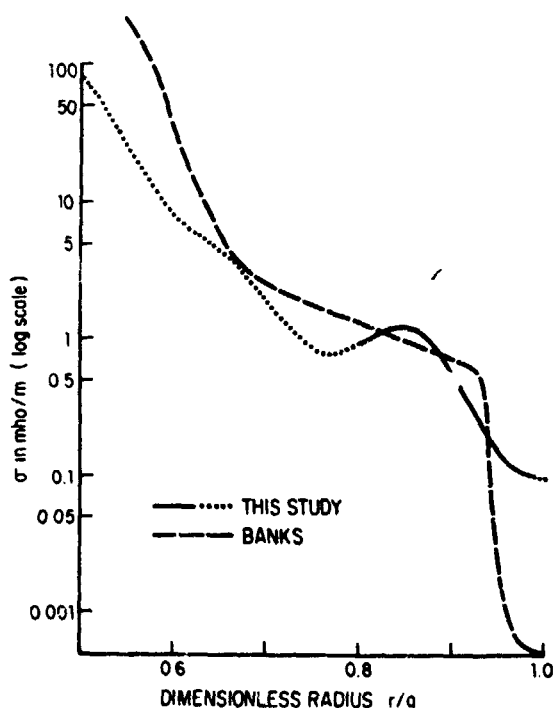


Figure 3. — Theoretical conductivity model used to fit the data of figure 1.

theoretical model that fits the observations to within about one standard error: the continuous line in figure 2 shows the response of the model, which is shown in figure 3. Such a simple curve is not an adequate description of the model, and the ambiguities due to lack of resolution and lack of precision are of considerable importance.

decomposed into spherical harmonics. It has been found that the driving field consists largely of an $\ell = 1$ harmonic (a uniform field) at the longer periods, and a physical mechanism is available to account for this observation.

To obtain information about the conductivity from 100 km to 1000 km it is necessary to record very slow variations in the field from about 1 cycle per day to 1 cycle per year. In this frequency band there are periodic and aperiodic sources, both of which can be used for inversion purposes. But the amplitudes of the variations are very small, often only a few parts in 10^5 when compared with the steady field, so that in practice the application of cross correlation and other statistical techniques is vital.

The theory is made much easier if we assume the earth is spherical and that the conductivity varies only as a function of radius. This approximation is clearly rather poor near the surface, where there are oceans and other irregularly distributed conductors. By concentrating on the lower frequencies we hope to "see through" the surface layer and observe a region where the assumption is valid. With the further assumption that the source field is uniform, it turns out that the key observation is the ratio of the amplitudes of the vertical and northward horizontal fields considered as a function of frequency. This quantity should vary with position like the sine of the geomagnetic latitude; that it does for a number of stations [Banks, 1969] bears out the validity of the two assumptions. The geomagnetic data published by Banks were used in the application of the Backus-Gilbert theory to the electrical conductivity problem. These data cover the range 1 cycle per day to 1 cycle per year from which the frequency band 0.2 cycles per day to 0.01 cycles per day was extracted for purposes of the work under discussion. In this interval Banks gave 35 frequencies at which he estimated the amplitude ratio of the vertical to horizontal field (normalized by the sine of the latitude). Of course, this ratio has a phase as well as a magnitude, but the former had a much larger uncertainty and was not included in my analysis.

Application of Backus-Gilbert Method

Mathematical details required in finding the data-kernels for electrical conductivity, which is a nonlinear problem, are given by Parker [1970]. The 35 data described above and their standard errors are shown in figure 2. An iterative scheme not discussed in this article was used to generate a

A "tradeoff" contour diagram is shown in figure 4. Spread (the length over which the real conductivity is smoothed to obtain an estimate) is contoured against relative error and position in the earth. The D_2 criterion described earlier was used here. The diagram is arbitrarily terminated at the top since errors larger than 100% are not very interesting, but the bottom of the diagram is the line of minimum error: no matter what resolution we may tolerate this is the smallest error possible. If the relative error level of 0.2 or 20% is traced across the diagram, it will be seen that the spread is unreasonably large until $r/a = 0.8$ or so. Below this depth our estimate is a virtually meaningless average over the whole mantle, made even more valueless by the fact that the linearization approximation breaks down for such large ranges of σ . This simply shows that we know nothing about conductivity below $r/a = 0.8$.

Continuing the 20% level across, we can see that resolution lengths improve to $0.07a$ or so for r/a between 0.85 and 0.95. If we would like a more precise estimate than one with 20% error, we must take a larger averaging length as shown.

The purpose here is to demonstrate how the nonuniqueness information generated by the Backus-Gilbert technique is used. A geophysically interesting question concerns the appearance of a conductivity minimum at $r/a = 0.75$. The tradeoff diagram shows that the minimum is not a believable feature of the model. In a case like this, the errors ascribed to a solution are as important as the solution itself.

Discussion

In recent years our inverse problem has received some ingenious analytical attention: *Bailey [1970]* gives a direct solution based on the application of Krönig-Kramers relations to the Riccati equation, and *Weidelt [1970]* achieves the same objective with the *Gelfand and Levitan [1951]* procedure of quantum mechanics. Both techniques assume the availability of data that are perfectly accurate and cover all frequencies. Before these results were known, of course, iterative model building was used.

I have given here a very brief sketch of the application of the Backus-Gilbert approach. I hope I have made it plain why it is necessary to go beyond the production of a model fitting the data to a description of the uncertainties: I have indicated the form of the description in terms of a trade-off diagram, and given examples of its use. The Backus-Gilbert method is a general theory and I hope we shall see a wider use of it in the future.

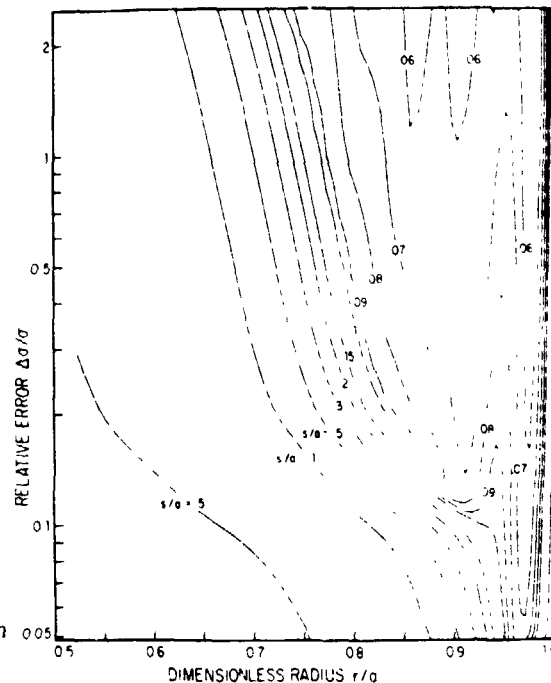


Figure 4. "Trade-off" contour model. Spread is plotted against relative error and position of Earth.

DISCUSSION

Newton: You seem to be implying that the main limitation on the resolving power comes from the number of observations. There is another problem that has to do with the linear dependence of the observations. Your example is a vivid example of just that. It is implied by the theory that two observations are not going to give you any more than a single observation. In another case you may not be able to see that so easily.

Parker: This is very true. But, if you go through the procedure carefully it comes out automatically.

Unidentified speaker: Could you use this theory to plan your experiment?

Parker: Yes. It can be used in that way, especially in a problem like the conductivity problem I discussed. Another case is satellite planning.

Lax: You mentioned data of the order of 500 or 1000. In that case there are some numerical problems in minimizing a quadratic form under a linear restriction. Can you comment on that?

Parker: This is a very important problem. The problem of linearizing very large quadratic forms is exactly the same problem as diagonalizing a large matrix. We try to overcome the difficulties by examining the spectrum of the eigenvalues you get. This is feasible.

Sondhi: How do the results depend on the criterion you use for the concentration of the linear combination of delta functions?

Parker: This has not really been explored.

Sondhi: Is there any difficulty in extending this when the G 's are known only at discrete values of r ? You have a certain number of kernel functions. What if you know the values of the samples of the kernel function only, and not the continuous function?

Parker: I think one could formulate it that way. We haven't in the geophysical sciences.

King: I think you've shown beautifully the limits of information retrieval using linear inversion techniques. This is basically due to the fact that there is a spread in your weighting function that no amount of linear gamesmanship in terms of manipulation will surmount. This same restriction does not seem to apply to nonlinear combinations of the observations. I think there is a limit in principle.

Parker: I think you are quite correct. If the inverse problem is linear you will gain nothing by using nonlinear averages. If, however, the inverse problem is nonlinear there must be available deeper insight using nonlinear theory.

CULLING OF REDUNDANT DATA

Freeman Gilbert

University of California at San Diego

1 N73-11620

This paper considers some practical numerical matters [Gilbert, 1971]. This is in effect an extended answer to the question, "What do you do when you have a large data set?" We first demonstrate the inversion of a large data set, with its errors, and then discuss the tradeoff curve, or resolving power calculations. We consider inverse problems of the earth, represented in the idealized form of a single scalar function of a single coordinate $m_E(x)$. A model of the earth that may or may not fit the data is $m(x)$. For each model, we calculate a gross earth functional g_i , which is some observable quantity that depends on the whole earth or some fraction of the whole earth; γ_i is the observed value of g_i , the outcome of some experiment or set of experiments, and G_i is the Frechet' kernel, or the first derivative, which permits the effect δg_i of a small change in the model δm to be computed:

$$\int_0^1 \delta m(x) G_i(x) dx = \delta g_i \quad (1)$$

In practice, almost all geophysical data are not Frechet' derivatives but ordinary differentials, and we consider only those.

In formulating inverse problems, we want to turn this relationship around and think of δg_i as being the difference between the functional for the real earth and the functional for some model. We may regard the model as a starting model or as a model in some stages of an iterative procedure, so that the functional for the real earth is the observed value of γ_i . It is the difference between the observed value γ_i and the calculated value of g_i that we call δg_i .

$$\begin{aligned} \delta g_i &= g_i(m_E) - g_i(m) \\ &= \gamma_i - g_i(m) \end{aligned} \quad (2)$$

We know the model, we don't know the real earth, and we seek to estimate this difference. It is clear that we have linearized the statement of the problem, and because we only have a finite number of data that represent linearized moments and the difference $\delta m(x)$, the answer cannot possibly be unique except under very pathological circumstances. This is both good and bad. It is bad because we will never know what the real earth is really like, but it is good because we can explore this line of uniqueness to find a model, any model, at the present time.

For example, we can regard the linear relationship between δg and δm for all the data as being a side condition, or constraint, and make the demand that we want to find the δm that is smallest in an rms sense:

$$\min \int_0^1 [\delta m(x)]^2 dx \quad (3)$$

We wish to minimize this quadratic form, and use those linear relationships between the differences in the data and the differences in the model as side conditions. This is a classical isoperimetric problem in the calculus of variations, and has a very simple solution. For this particular minimum condition, the Lagrange-Euler equations turn out to be algebraic, which is always nice.

Two other criteria may be used: the flattest model of perturbation that fits the data

$$\min \int_0^1 [\delta m'(x)]^2 dx \quad (4)$$

or the smoothest model of perturbation that fits the data

$$\min \int_0^1 [\delta m''(x)]^2 dx \quad (5)$$

From an esthetic point of view we find that the criterion (5) is the one that gives us the nicest results, but any one can be chosen. If we use the first criterion, for the smallest rms δm , the result of solving the Lagrange-Euler equation is:

$$\delta m(x) = \sum_j v_j G_j(x) \quad (6)$$

where the v are the Lagrange multipliers we put in to handle the side conditions. We find in this case that the perturbation is just a linear combination of the Frechét kernels. This may be substituted back into Eq. (1) to give:

$$\sum_j \int_0^1 G_i(x) G_j(x) v_j dx \quad (7)$$

So far, I have assumed that there are no errors in the data. The kernel $G_i G_j$ is a symmetric matrix that is at least positive semidefinite, and is positive definite if you are lucky. The expression in Eq. (7) is an inner product of a very simple kind. Thus, these values of v_j may be found by solving the system of linear equations. Equation (6) allows construction of δm , which when added to the starting model gives a new model whose properties, as predicted by perturbation theory, agree with the properties measured.

Of course, since the inverse problem is not linear, this process has to be iterated, and we have found that when you start one of these iterative procedures it either blows up instantly or converges very rapidly.

We now consider the effect of errors in the data. Suppose each datum has a standard error σ , and we require only that the data be fitted to within some band, say one or two standard errors. Once again, the isoperimetric problem may be set up, with the side conditions

$$\delta g_i - \sigma_i \leq \int_0^1 \delta m(x) G_i(x) dx \leq \delta g_i + \sigma_i \quad (8)$$

and minimize the integral of $(\delta m)^2$. Once again δm turns out to be a linear combination of Frechét kernels, and we have to solve

$$\delta g_i - \sigma_i \leq \sum_j \int_0^1 G_i(x) G_j(x) v_j dx \leq \delta g_i + \sigma_i \quad (9)$$

Now it is more difficult to solve these inequalities than it is to solve equations, and usually some kind of linear programming technique is used. Suppose, however, that we are very fortunate, and in addition to having a matrix that is not only symmetric and positive definite, we have a diagonal matrix. Then solving Eq. (9) is very easy. In this case, when the G is orthonormal, the integral of $(\delta m)^2$ is easily shown to be just the sum of the Lagrange multipliers

$$\int_0^1 (\delta m)^2 dx = \sum_i v_i^2 \quad (10)$$

To minimize Eq. (10), we need only make each of the Lagrange multipliers as small as possible subject to the side conditions, which are now

$$v = \delta g_i - \sigma_i \leq v_i \leq \delta g_i + \sigma_i = u \quad (11)$$

This is done merely by taking $v = v$ if $v > 0$, $v = u$ if $u < 0$, and $v = 0$ otherwise.

Because of the ease of solving these inequalities, it is evident that we need to diagonalize the matrix

$$G_{ij} = \int_0^1 G_i(x) G_j(x) dx \quad (12)$$

At the same time we should allow for the possibility that each datum may not have a single standard error associated with it. There may be interactive errors . . . covariances. In addition to diagonalizing the inner product matrix G we need to diagonalize the covariance matrix V .

The matrix G will always be symmetric positive semidefinite; V will always be symmetric positive semidefinite, and will be positive definite if the experiments are properly conducted. Our problem is the classical one of diagonalizing simultaneously two quadratic forms, both of which are symmetric, and one of which is positive definite. This can readily be done [Whittaker, 1937, pp. 178-183].

Let T be the diagonalization transformation

$$\begin{aligned} T \cdot V \cdot T^T &= E \\ T \cdot G \cdot T^T &= I \end{aligned} \quad (13)$$

where E is diagonal, and I is the diagonal unit matrix. The new gross earth functionals and the new observables are transformations of the old ones

$$f = T \cdot g \quad (14)$$

$$\phi = T \cdot \gamma$$

and new Frechet kernels F are generated by linear transformation of the old ones

$$F(x) = T \cdot G(x) \quad (15)$$

The standard errors of the now independent data are just the square roots of the diagonal entries of the diagonal variance matrix

$$S_i = E_{jj}^{1/2} \quad (16)$$

The relationship between the perturbation of the model and the perturbation of the data kernels is now given in terms of the new Frechet kernels by

$$S'_{fi} = \phi_i - f_i = \int_0^1 \delta m(x) F_i(x) dx \quad (17)$$

where

$$\int_0^1 F_i(x) F_j(x) dx = S_{ij} \quad (18)$$

We proceed to the construction of T in two steps. First, we find the orthogonal matrix Q which diagonalizes the covariance matrix V .

$$V = Q \cdot U \cdot Q^T \quad (19)$$

In most cases U will already be diagonal, when there are no covariances. A transformed inner product matrix H is then

$$H = Q \cdot G \cdot Q^T \quad (20)$$

Second, we start with the quantities v_i , which are the square roots of the errors

$$v_i^2 = U_{ii} \quad (21)$$

and form a new symmetric matrix A by dividing the rows and columns of H by the v

$$A_{ij} = v_i^{-1} H_{ij} v_j^{-1} \quad (22)$$

The matrix A may now be diagonalized by an orthogonal matrix R

$$\Lambda = R \cdot A \cdot R^T \quad (23)$$

The standard errors of the new data are just the square roots of the reciprocals of Λ . Now if we put P the following nondiagonal matrix

$$P_{ij} = S_i R_{ij} v_j^{-1}$$

then the transformation that diagonalizes G and V simultaneously is

$$T = P \cdot Q$$

The new error matrix is just

$$E = P \cdot U \cdot P^T = \Lambda^{-1}$$

The problem is now reduced to finding the two orthogonal matrices Q and R , which in most cases amounts only to finding R . The procedure we use for diagonalizing these matrices is essentially a classical, stable method, due to Householder [Wilkins, 1965, pp 290-332]. We will return to that after considering the problem of tradeoff curves.

After doing the construction problem—creating a model that fits the data by means of the ideas given above—one presumably needs to do resolving power calculations, as discussed in the preceding paper by Parker. We consider his matrix for the "spread" S and the covariance matrix V . Then W is a weighted sum of these matrices

$$W = S \cos \theta + V \sin \theta$$

where a tradeoff curve is generated as θ runs from 0 to $\pi/2$. Generally, we want to solve some system of linear equations like

$$W \cdot A = U$$

Now if both S and V were diagonal, the equations could be easily solved. Since they are both positive semi-definite, we can go through the same thing again, and diagonalize them, and turn this into a system of independent equations.

In practice, we find the orthogonal transformations which diagonalize the matrices as follows. A matrix such as A in Eq. (23) is first reduced to tridiagonal form by Householder's method. The eigenvalues of the tridiagonal matrix are the eigenvalues of A and can be found by Sturm's bisection method. Finally, the eigenvectors of the

tridiagonal matrix are found by *Wilkinson's* [1965] method of iterative improvement, and are transformed into the eigenvectors of *A* by the tridiagonalization transformation. We have found this approach to be extremely stable and rapid.

Now, to solve a system of symmetric linear equations like this, about $(1/6)n^3$ operations are needed. If *p* points are required to generate a tradeoff curve, then $(1/6)n^3p + O(n^2p)$ operations are needed. In the diagonalization method, one needs $(2/3)n^3$ operations to tridiagonalize and another $(2/3)n^3$ to get the eigenvalues. That is done once for the entire tradeoff curve, and the number of points on the tradeoff curve gives a further $O(n^2p)$ operation. If *p* is large, it is possible that the number of operations required to solve the system of linear equations will be larger than the cost of diagonalizing once and for all for each tradeoff curve and then doing n^2 operations. For *n* greater than *p*—that is, when the data set is larger than the number of points needed on the tradeoff curve—then the ratio of operations is *p*/8 (in our applications *p* is about 25 and *n* is 100 or more). In other words, if there are more than eight points on the tradeoff curve it is more efficient to diagonalize first instead of solving the system of linear equations. That is for a single tradeoff curve, so it is always more efficient to diagonalize.

If the smoothness criterion is the *D*₂ criterion discussed by Parker, *S* must be rediagonalized for every *r*₀ changes. If the *D*₁ criterion is used, no *r*₀ appears, and the matrix *S* takes the same form for all radii, so diagonalization is done once, not only for each tradeoff curve, but for all of them. If there are *C* curves then the diagonalization method is *pC*/8 more efficient with the use of the *D*₁ criterion. If, for example, there are 24 points on a curve and 50 curves, it is 150 times more efficient, which is not trivial financially. To do the equivalent of 300 data, with 24 points per curve and 50 curves, takes about 410 sec. on a CDC 6600, which is not expensive. This kind of calculation is not done very often.

The actual procedure we use is to make the model fit the data after diagonalizing everything—the constructive part of the problem. The data are culled, by accepting only those relative standard errors that are less than 100%. In most geophysical problems, the effect is to reduce the amount of data from about 300 to about 30. We have termed this *redundancy*. But those 30 represent the projection of the contributions of all of the 300 data into that good data set. This turns out to be an extremely efficient and rapid procedure. Typically, when a model agrees with the data within a few percent, and we wish to get it to agree within 0.2%, two iterations are sufficient for 300 data.

In doing the tradeoff curves, we do exactly the same thing. What the large errors do, after diagonalization of these 300 forms, is essentially allow you, at an extreme price, to narrow down the averaging width a little. The extreme price is enormous errors. In a practical case all the errors greater than 100% are discarded and we forget about the top part of the tradeoff curve.

This is a very simple, practical scheme. We have spent a lot of time coding these matrix operations for efficiency. I followed all the recommendations that *Wilkinson* has made in his book on the algebraic eigenvalue problem, plus one or two of our own. We store the matrix in linear, diagonal form, and operate successively from one column to the next. So, in fact, so far as the computer is concerned, we start with cell 1, go through sequentially, and wind up with cell number $n(n+1)/2$. Computers like this, so the efficiency goes up enormously.

DISCUSSION

Falcone: What is the form of the Frechét kernel G_f ?

Gilbert: It is a function of radius that may have one or two step discontinuities. Of course, it is represented as a set of points for machine calculations. Normally it is a linear combination of squares or cross products of eigenfunctions.

Unidentified speaker: Did you say you could diagonalize a positive semidefinite matrix?

Gilbert: Actually, from the point of view of the computer, these matrices are extremely singular. What is done, with the Sturm bisection method, is to find the largest eigenvalue, the next largest, and so on, until one with value 1 or less occurs. Since this implies a relative error of 100%, the calculation is terminated.

Unidentified speaker: You implied that you could diagonalize a matrix in a finite number of steps. Doesn't finding the eigenvalues, the roots of a polynomial, take an infinite number of steps?

Gilbert: In practice, of course, we quit when the relative value is known to 10^{-14} . That figure of $(2/3)n^3$ referred to the number of steps needed to tridagonalize. The bisection method then completes the process rapidly.

Moses: Could this be used for Fourier transformation? You might have data with errors, for which you seek the Fourier transform.

Gilbert: Yes. As a matter of fact, you could certainly look at the problem of digital Fourier analysis as being the inverse problem of trying to find the spectrum, given the time function.

Thellman: Aside from the question of the size of the standard errors, we found that near-zero eigenvalues were associated with too much covariance between points in the model, which implies too much resolution.

Gilbert: If you go ahead and diagonalize the covariance matrix by a simple orthogonal transformation, you are projecting into a new orthogonal space, where your data can be ranked according to standard error. If you have too many data because you made too many overlapping measurements, you make this projection and look at the standard error. If it begins to build up, you just cut off the data set.

Unidentified speaker: Adding on data that have large standard errors should not make the errors wider in the model. It might make the mathematics easier if you throw it away.

Gilbert: That is quite true. There is nothing wrong with having a completely redundant set of data. In fact, one of the easiest ways in a big machine program to substitute a new datum for an old datum is to just read it in, because the old one automatically gets shoved down in the eigenvalue stack. We have a redundancy factor of something like 10 or 15—with 300 data, we come up with 20 or 30 significant data.

APPLICATION OF A STOCHASTIC INVERSE TO THE GEOPHYSICAL INVERSE PROBLEM

Thomas H. Jordan
and
J. Bernard Minster

California Institute of Technology

N73-11621

ABSTRACT

The inverse problem for gross earth data can be reduced to an underdetermined linear system of integral equations of the first kind. Discussed in this paper is a theory for computing particular solutions to this linear system based on the stochastic inverse theory presented by Franklin. The stochastic inverse is derived and related to the generalized inverse of Penrose and Moore. A Backus-Gilbert type tradeoff curve is constructed for the problem of estimating the solution to the linear system in the presence of noise. It is shown that the stochastic inverse represents an optimal point on this tradeoff curve. In the appendix a useful form of the solution autocorrelation operator as a member of a one-parameter family of smoothing operators is derived.

INTRODUCTION

This paper addresses the problem of computing particular solutions to certain finite systems of linear integral equations that arise in the study of the earth's interior. On the earth's surface, quantities can be measured that are functionals of the distribution of physical parameters at depth. The earth's mass and moments of inertia, functionals of the density distribution, and the travel times of seismic rays, functionals of the velocity distribution, are examples of this kind of data. In essence, the geophysical inverse problem is this: given the observations of a finite number of these functionals, what is the distribution in the earth's interior of the parameters on which they depend? The formulation of the problem used in this paper relates small changes in a chosen model to small changes in the data through a linear transformation. With the use of this first-order approximation the problem is reduced to the solution of an underdetermined system of linear integral equations. Since observations are invariably contaminated by errors introduced through inaccurate measurement, only *estimates* of the values of the data functionals for earth are available, and estimates of the solution must be made with this realization. A general and extensive theory concerning the solution of the linear geophysical inverse problem for inaccurately known data has been provided by *Backus and Gilbert* [1967, 1968, 1970]. We draw heavily from the concepts and terminology developed by these authors. Our concern here is the application of stochastic inverse theory presented by *Franklin* [1970] to obtain particular solutions to the linear system. This paper represents an expansion on the work of *Jordan and Franklin* [1971]. Since the geophysical inverse problem is an example of a canonical problem in the mathematics of profile inversion, the discussion is kept general to facilitate other applications. Numerical results from the inversion of gross earth data are the subject of a separate paper [*Jordan and Anderson*, in preparation] and are not presented here.

A *model* is here defined as a real-valued function m specified on the closed interval $[0,1]$. In the geophysical problem m represents the distribution of, say, density as a function of radius in the earth. The model space \mathcal{M} is the Hilbert space completed from the space of real-valued piecewise continuous functions square integrable on the interval $[0,1]$. The inner product of a pair of functions $m, m' \in \mathcal{M}$ is written $m \cdot m'$ and taken equal to the scalar value of the integral $\int_0^1 m(x) m'(x) \rho(x) dx$. The weight function $\rho(x)$ is required to be strictly positive on $[0,1]$. Associated with this inner product is the norm $\|m\| = (m \cdot m)^{1/2}$. Given a linear operator L , mapping

M into itself and an element m , the vector $L \cdot m$ in M whose value at each point x equals the integral

$\int_0^1 L(x, x') m(x') \rho(x') dx'$ can be computed. The function $L(x, x')$ is the kernel of the operator L . The kernel of the identity operator I is $\delta(x - x')/\rho(x')$ where the numerator is the Dirac delta distribution. L^* denotes the transpose of L . If $L = L^*$, L is symmetric.

A *linear data functional* is a linear functional defined over the model space M . (If a function is nonlinear but Fréchet differentiable, then it can be approximated by a linear data functional at each point.) The inverse problem is to estimate the function m_0 , representing the actual density distribution in the earth, given the observed values of an ordered set $D^N = \{d_i(m) : i = 1, 2, \dots, N\}$ of N linear data functionals, say the earth's mass and moment of ($N = 2$). Corresponding to the functional $d_i(m)$ there exists a unique $a_i \in M$, the kernel of d_i , such that

$$a_i \cdot m = d_i(m) \quad i = 1, 2, \dots, N \quad (1)$$

Each $d_i(m)$ can be considered as the i th component of a vector d in the N -dimensional Euclidian space E^N . The inner product between two vectors d and d' in E^N is written dd' and equals

$$\sum_{n=1}^N d_i d_i'$$

The inner product is assumed to be dimensionally homogeneous. Let d_0 be the vector whose components are the observed data: then, if the data are error-free, m_0 satisfies the operator equation

$$A \cdot m_0 = d_0 \quad (2)$$

where $A : M \rightarrow E^N$ is the operator whose i th row is the vector a_i mapping M onto the subspace $R(A) \subseteq E^N$, the range space of A . Equation (2) implicitly assumes that $d_0 \in R(A)$.

THE GENERALIZED INVERSE

For finite N the problem of computing the solution to Eq. (2) is ill-posed in the sense that the solution is not unique. In fact, A possesses a null manifold $N(A)$ of infinite dimension. For each $h \in N$, $A \cdot h = 0$. If $h_n : n = 1, 2, 3, \dots$ is basis of $N(A)$, then the general solution to Eq. (2) can be written

$$m = \bar{m} + \sum_{n=1}^{\infty} \alpha_n h_n \quad (3)$$

where \bar{m} is any particular solution to Eq. (2) and the coefficients α_n are arbitrary real numbers. One particular solution of interest is given by the generalized inverse of A [Moore, 1920; Bjerhammar, 1952; Penrose, 1955;

Tseng, 1949]. $R(A^*)$ is the space spanned by the set $\{a_i : i = 1, 2, \dots, N\}$, so $M = N(A) \oplus R(A^*)$. The operator A has a unique generalized inverse $A^\dagger : E^N \rightarrow M$ such that

$$A \cdot A^\dagger = P_{R(A)}, \quad A^\dagger A = P_{R(A^*)} \quad (4)$$

where $P_{R(A)} : E^N \rightarrow R(A)$ and $P_{R(A^*)} : M \rightarrow R(A^*)$ are orthogonal projection operators (that is, $P \cdot P = P$ and $P^* = P$). The estimate

$$\bar{m} = A^\dagger d_0 \quad (5)$$

is the unique solution that minimizes the norm $\|m\|$. Using Eqs (2) and (4) we write

$$\bar{m} = A^\dagger A \cdot m_0 = P_{R(A^*)} \cdot m_0 \quad (6)$$

Thus, this solution corresponds to the orthogonal projection of any solution onto the subspace $R(A^*)$. It is easily shown that $A^\dagger = A^*(A \cdot A^*)^\dagger$, reducing the computation of A^\dagger to the determination of the generalized inverse of a symmetric operator on E^N . If A has rank N , then $(A \cdot A^*)^\dagger = (A \cdot A^*)^{-1}$ since $A \cdot A^*$ is nonsingular. If A has rank less than N , other algorithms for computing $(A \cdot A^*)^\dagger$ can be used, such as that of *Ben-Israel and Charnes*, [1963].

The form of Eq. (6) illustrates an important point: since the data kernels do not form a complete set, the value $m_0(x)$ cannot be determined at each $x \in [0,1]$; rather a linear average of $m_0(x)$ given by

$$\bar{m}(x) = \int_0^1 P_{R(A^*)}(x, x') m_0(x') \rho(x') dx' \quad (7)$$

is obtained. Usually, it is desirable to compute an average which is at least *localized* at each point. Roughly speaking, an average is said to be localized if the contributions to the integral (7) are small away from the point x . The concept of the localized linear average is of central importance in the Backus-Gilbert theory for the resolving power of a linear system [Backus and Gilbert, 1968]. In practice, these considerations can limit the usefulness of the generalized inverse. Backus and Gilbert [1968] examined the kernel of $P_{R(A^*)}$ and found that, for typical sets of eigenfrequency data, the linear averaging associated with this projection was not sufficiently localized to provide useful estimates of the density in the earth. Furthermore, Eq. (5) was derived under the assumption that the data are perfectly well-known. Actually d_0 is only an estimate of the vector $A \cdot m_0$ that has been corrupted by error or "noise" entering through observational errors and computational inaccuracies. Neglecting this error can yield model estimates with large statistical uncertainties [Backus and Gilbert, 1970].

Nevertheless, solution of the perturbation equations by the minimization of a specified norm is an appealing method for selecting particular solutions to Eq. (2). For instance, if on the basis of auxiliary information there is good reason for believing that the representation m_0 is somehow close to a particular starting model, such a minimization can provide the smallest perturbation available—in a norm sense—that satisfies Eq. (2). The limitations of the generalized inverse can be overcome by appealing to a stochastic formulation of this linear problem [Franklin, 1970].

THE STOCHASTIC INVERSE

The equation corresponding to Eq. (2) for inaccurate data can be cast in the form

$$A \cdot m_0 + n = d_0 \quad (8)$$

where $n \in E^N$ is a vector containing the noise components. Since these components have some unknown scalar value, the error is described only in terms of its statistics. Treated here is the case for Gaussian noise with zero mean. Following Franklin [1970] we consider Eq. (8) to be a sample of the stochastic equation

$$A \cdot p_s + p_n = p_d \quad (9)$$

where p_s is a stochastic process describing the solution and is defined over M , and p_n is the noise process and p_d the data process, both defined over E^N . The problem is to construct the best linear unbiased estimate \bar{p}_s of the solution process p_s resulting from the application of some linear operator B to the data process p_d :

$$\bar{p}_s = B p_d \quad (10)$$

To ensure that this estimate is unbiased, p_s and p_n must have zero expectation.

The process p_s has zero expectation if the expected value of the random variable $p_s \cdot g$, denoted $E[p_s \cdot g]$, is zero for all g in the model space. So long as convergence in the quadratic sense is sufficient, then p_s can be represented by the decomposition:

$$\sum_{n=1}^{\infty} a_n f_n$$

where $\{f_n : n = 1, 2, 3, \dots\}$ is some orthonormal basis for M and $\{a_n : n = 1, 2, 3, \dots\}$ is a set of orthogonal Gaussian random variables. The autocorrelation operator of the process p_s is the symmetric operator C_{ss} mapping the model space into itself such that $g \cdot C_{ss} \cdot g = E[(p_s \cdot g)^2]$ for all $g \in M$. The term on the right-hand side is the variance of the random variable $p_s \cdot g$; therefore, C_{ss} is positive semidefinite. By the Karhunen-Loève theorem [Loève, 1955, p. 478], $\{f_n\}$ are the eigenfunctions of C_{ss} (taken here to be normalized) and $\{\alpha_n^2 : n = 1, 2, 3, \dots\}$, the variances of the a_n , are its eigenvalues

$$C_{ss} = \sum_{n=1}^{\infty} \alpha_n^2 f_n f_n \quad (11)$$

The linear operator $f_n f_n$ is defined by $(f_n \cdot f_n) \cdot g = f_n(f_n \cdot g)$. If the spectral coefficients α_n^2 equal unity for all n , then Eq. (11) reduces to the completeness relation

$$1 = \sum_{n=1}^M f_n f_n \quad (12)$$

Correspondingly, the solution-data cross-correlation operator $C_{sd} : E^N \rightarrow M$ is defined by the bilinear form $E[(\sim \cdot p_s)(p_d \sim)]$. Evidently, $C_{sd} = C_{ds}^*$. The following expansions are obtained for the data autocorrelation operator and the solution-data cross-correlation operator from Eq. (9)

$$\begin{aligned} C_{dd} &= A \cdot C_{ss} \cdot A^* + A \cdot C_{sn} + C_{ns} \cdot A^* + C_{nn} \\ C_{sd} &= C_{ss} \cdot A^* + C_{sn} \end{aligned}$$

Assuming that the solution and noise processes are uncorrelated, $C_{sn} = C_{ns} = 0$, and the above expressions reduce to

$$\begin{aligned} C_{dd} &= A \cdot C_{ss} \cdot A^* + C_{nn} \\ C_{sd} &= C_{ss} \cdot A \end{aligned} \quad (13)$$

In this form, the data autocorrelation operator C_{dd} is nonsingular if C_{nn} is positive definite.

The best linear estimate of the solution process p_s is the process \bar{p}_s which minimizes the quadratic form $\epsilon^2(g) = E\{[(p_s - \bar{p}_s) \cdot g]^2\}$ for all $g \in M$. Substitution using Eq. (10) and expansion of the autocorrelation operator of $p_s - \bar{p}_s$ yields

$$\epsilon^2(g) = g \cdot C_{ss} \cdot g - g \cdot C_{sd} B^* \cdot g - g \cdot B C_{ds} \cdot g + g \cdot B C_{dd} B^* \cdot g$$

The first and second variations of the functional ϵ^2 with respect to a variation of the vector $q = B^* \cdot g$ (g fixed) are

$$\begin{aligned} \delta(\epsilon^2) &= 2(q C_{dd} - g \cdot C_{sd}) \delta q \\ \delta^2(\epsilon^2) &= \delta q C_{dd} \delta q + 2(q C_{dd} - g \cdot C_{sd}) \delta q^2 \end{aligned} \quad (14)$$

The functional $\epsilon^2(g)$ is stationary if and only if $\delta(\epsilon^2) = 0$ for all arbitrary variations δq . Therefore, the linear combination $C_{dd}q - C_{ds} \cdot g$ is required to be zero for all $g \in M$. Assuming C_{dd} to be nonsingular, this is true if and only if $B = C_{sd} C_{dd}^{-1}$. For this choice the second variation reduces to the positive definite form $\delta^2(\epsilon^2) = \delta q C_{dd} \delta q$. Since this stationary point is unique and a minimum, the best linear estimate of p_s is

$$\bar{p}_s = C_{sd} C_{dd}^{-1} p_d \quad (15)$$

For the particular sample of the data process d_0 , the best estimate of m_0 is, using Eq. (13),

$$\bar{m} = C_{ss} \cdot A^* (A \cdot C_{ss} \cdot A^* + C_{nn})^{-1} d_0 \quad (16)$$

This stochastic inverse is well suited for numerical computations. Its validity is limited only by the assumptions that the solution and noise are uncorrelated and that C_{nn} is positive definite. In practice, it is commonly assumed that the error components are themselves uncorrelated; then

$$C_{nn} = \begin{bmatrix} \sigma_1^2 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \sigma_N^2 \end{bmatrix} \quad (17)$$

The positive real number σ_i^2 is just the variance of the i th datum. This matrix is positive definite, thus ensuring the existence of C_{dd}^{-1} .

Statistical information concerning p_s and p_n contained in an ensemble of samples can be used to estimate their autocorrelation operators. The variances in Eq. (17), for instance, can be estimated from the scatter in the data.

Information concerning the distribution of the solution and noise processes has been used in the solution of the linear estimation problems previously. Equation (16) is analogous to the results of Wiener's theory for construction of an optimum infinite-lag smoothing filter [Davenport and Root, 1958] and was explicitly obtained by Strand and Westwater [1968]. It reduces to Twomey's [1963] results for the special case $C_{ss} = I$, $C_{nn} = \gamma I$.

In the geophysical problem it occurs that no statistics for p_s exist. Then, what is the significance of characterizing the solution as a sample of a Gaussian process with a zero expectation and an autocorrelation operator C_{ss} ? To answer this question we assume that the noise is identically zero and that both C_{ss} and $A \cdot C_{ss} \cdot A^*$ are positive definite. Under these conditions the best estimate Eq. (16) reduces to

$$\bar{m} = C_{ss} \cdot A^* (A \cdot C_{ss} \cdot A^*)^{-1} d_0 \quad (18)$$

This solution has a simple geometrical interpretation. Let M' be the space whose elements are the members of M with an inner product defined as the positive definite bilinear form

$$m \cdot_L m' = \int_0^1 \int_0^1 L(x, x') m(x) m(x') \rho(x) \rho(x') dx' dx \quad (19)$$

$m, m' \in M'$, where $L = C_{ss}^{-1}$. The solution autocorrelation operator C_{ss} is idempotent with respect to this product ($C_{ss} \cdot_L C_{ss} = C_{ss}$); it is the identity operator on M' . Define $A': M' \rightarrow E^N$ as the linear operator $A \cdot C_{ss}$. The elements of $N(A')$ are the same as the elements of $N(A)$. Substituting $A' \cdot_L m_0$ for d_0 in Eq. (18) we obtain

$$\bar{m} = P_{R(A'*)} \cdot_L m_0 \quad (20)$$

where the operator $P_{R(A'*)} : M' \rightarrow R(A'*)$ is the orthogonal projection operator onto the range space of A'^* . In this notation,

$$P_{R(A'*)} = A'^* (A' \cdot_L A'^*)^{-1} A' \quad (21)$$

Thus, the solution given by Eq. (18) is the symmetric projection of any solution, in particular m_0 , onto the manifold $R(A^*)$, in the inner product being defined by Eq. (19) in terms of the solution autocorrelation operator.

Often it is useful to prescribe C_{SS} as a smoothing operator corresponding to a priori assumptions about the smoothness of the solution. This is a way to incorporate additional information about the solution not contained in the data set D^N . Because this point is important for physical applications, it bears further discussion. To illustrate what it means for C_{SS} to be a smoothing operator or, in other terminology, a low-pass filter, we return to its Karhunen-Loeve expansion Eq. (11). Assume that the orthonormal basis $\{f_n : n = 1, 2, 3, \dots\}$ has been ordered such that, if f_i is a "smoother" function than f_j —say, has fewer zero crossings in $[0,1]$ —then $i < j$. In the appendix we consider an ordered complete set with this property provided by the eigenfunctions of a Sturm-Liouville system. We will call C_{SS} a smoothing operator if $\alpha_i^2 > \alpha_j^2$ for $i < j$. This allows an ordering to be defined on M : a vector m_1 is said to be smoother than a vector m_2 if

$$\frac{\|m_1\|_L}{\|m_1\|} < \frac{\|m_2\|_L}{\|m_2\|}, \quad L = C_{SS}^{-1} \quad (22)$$

Here, $\|m\|_L = (m \cdot m)^{1/2}$. Taking for m_1 the solution (18) and for m_2 the solution (5), we have, by the properties of projection, $\|m_1\| > \|m_2\|$ and $\|m_1\|_L < \|m_2\|_L$ if $m_1 \neq m_2$. Therefore, by Eq. (22), m_1 is smoother than m_2 is smoother than m_2 .

Also, numerical computations with sets of gross earth data have shown that the introduction of a smoothing operator for the signal autocorrelation operator can have the effect of providing more localized linear averages of the representation m_0 than those given by the integral (7).

THE TRADEOFF CURVE

The Backus-Gilbert theory of linear estimation [Backus and Gilbert, 1968, 1970] suggests that for the problem described in the previous section—the estimation of a function m_0 given the values of a set of linear data functionals d_0 corrupted by a random noise—there exists a tradeoff between the ability to resolve the detail of m_0 and the reliability of the estimate \hat{m} . In this section, we investigate the optimality of the stochastic solution (16) in terms of its relative position on a Backus-Gilbert type tradeoff curve. The generalized inner product $\sim^L \sim$ and norm $\|\sim\|_L$ introduced in the previous section will be retained throughout the analysis, but the use of primes to distinguish products is suspended by defining the generalized product over M instead of M' .

An estimate \hat{m} of the function m_0 satisfying

$$A^L m_0 + n = d_0 \quad (23)$$

is sought given A , d_0 and the statistics of a Gaussian noise process p_N from which n is a sample. The process p_N is assumed to have zero expectation and a positive definite autocorrelation operator C_{NN} . The null space $N(A)$ is populated by those members of h for which $A^L h = 0$. Since the data contain no information about the components of m_0 in the null space, the estimate \hat{m} is required to belong to the range space $R(A^*)$. This statement is equivalent to constraining \hat{m} to be a linear combination of the data kernels

$$\hat{m} = A^* b, \quad \text{for some } b \in E^N \quad (24)$$

The vector \mathbf{b} is to be determined by minimization of an appropriate scalar measure of the error of estimation of \mathbf{m}_0 , which we are free to construct.

One obvious measure of the error of estimation of \mathbf{m}_0 is the norm of the difference between \mathbf{m}_0 and $\bar{\mathbf{m}}$:

$$\epsilon_1^2(\mathbf{b}) = \|\mathbf{m}_0 - \mathbf{A}^* \mathbf{b}\|_{\mathbf{L}}^2 \quad (25)$$

By the orthogonality of the spaces $N(\mathbf{A})$ and $R(\mathbf{A}^*)$ it is clear that the projection of \mathbf{m}_0 in the null space contributes to ϵ_1^2 its full squared norm regardless of the choice of \mathbf{b} . In fact, ϵ_1^2 is minimized at the value $\|\mathbf{P}_{N(\mathbf{A})}^{\mathbf{L}} \mathbf{m}_0\|_{\mathbf{L}}^2$ for $\mathbf{b} = (\mathbf{A}^{\mathbf{L}} \mathbf{A}^*)^{\dagger} \mathbf{d}_0$. This is the solution given by the generalized inverse Eq. (18).

If the data were perfectly accurate, the best linear estimate of the function \mathbf{m}_0 would result from the application of the generalized inverse of \mathbf{A} to the data sample vector \mathbf{d}_0 . However, $n \neq 0$ implies an uncertainty in \mathbf{d}_0 and, correspondingly, in $\bar{\mathbf{m}}$. A measure of the uncertainty of any estimate of the form Eq. (24) due to noise in the data is the variance ϵ_2^2 of the projection of \mathbf{p}_n onto \mathbf{b} . By definition,

$$\epsilon_2^2(\mathbf{b}) = \mathbf{b} \mathbf{C}_{nn} \mathbf{b} \quad (26)$$

An attempt to minimize this error with respect to a variation of \mathbf{b} yields the trivial solution $\mathbf{b} = 0$.

In general, the two measures of error ϵ_1^2 and ϵ_2^2 compete: minimizing ϵ_1^2 alone results in an estimate for which ϵ_2^2 is large, while minimizing ϵ_2^2 without regard for how well the estimate satisfies Eq. (23) yields ϵ_1^2 at its maximum value. To explore the possibilities for a compromise, consider the quadratic measure of error

$$\epsilon^2(\theta, \mathbf{b}) = \epsilon_1^2(\mathbf{b}) \cos \theta + \epsilon_2^2(\mathbf{b}) \sin \theta \quad (27)$$

composed of a weighted sum of ϵ_1^2 and ϵ_2^2 . The weighting is parameterized by an angle θ that varies on the interval $[0, \pi/2]$, so that $\epsilon^2(0, \mathbf{b}) = \epsilon_1^2(\mathbf{b})$ and $\epsilon^2(\pi/2, \mathbf{b}) = \epsilon_2^2(\mathbf{b})$. For a given fixed θ , $\epsilon^2(\theta, \mathbf{b})$ can be minimized with respect to a variation of \mathbf{b} . The unique \mathbf{b} that minimizes $\epsilon^2(\theta, \mathbf{b})$ is

$$\mathbf{b}(\theta) = (\mathbf{A}^{\mathbf{L}} \mathbf{A}^* + \tan \theta \mathbf{C}_{nn})^{-1} \mathbf{A}^{\mathbf{L}} \mathbf{m}_0 \quad (28)$$

Replacing the vector $\mathbf{A}^{\mathbf{L}} \mathbf{m}_0$ by its estimate \mathbf{d}_0 and substituting Eq. (28) into (24) gives the best estimate of \mathbf{m}_0 as

$$\bar{\mathbf{m}}(\theta) = \mathbf{A}^* (\mathbf{A}^{\mathbf{L}} \mathbf{A}^* + \tan \theta \mathbf{C}_{nn})^{-1} \mathbf{d}_0 \quad (29)$$

Special cases of Eq. (29) include the generalized inverse ($\theta = 0$) and the stochastic inverse ($\theta = \pi/4$).

The solution $\mathbf{b}(\theta)$ can be put into Eqs. (25) and (26) to obtain ϵ_1^2 and ϵ_2^2 as functions of θ

$$\begin{aligned} \epsilon_1^2(\theta) &= \|\mathbf{m}_0\|_{\mathbf{L}}^2 + \|\mathbf{R}(\theta)\|_{\mathbf{L}}^2 - 2 \mathbf{m}_0^{\mathbf{L}} \mathbf{R}(\theta)^{\mathbf{L}} \mathbf{m}_0 \\ \epsilon_2^2(\theta) &= \mathbf{m}_0^{\mathbf{L}} \mathbf{A}^* \mathbf{Q}^{-1}(\theta) \mathbf{C}_{nn} \mathbf{Q}^{-1}(\theta) \mathbf{A}^{\mathbf{L}} \mathbf{m}_0 \end{aligned} \quad (30)$$

The operators $Q(\theta)$ and $R(\theta)$ appearing in these expressions are defined by

$$\begin{aligned} Q(\theta) &= A^L A^* + \tan \theta C_{nm} \\ R(\theta) &= A^* Q^{-1}(\theta) A \end{aligned} \quad (31)$$

Equations (30) determine a curve parameterized by θ in the $e_1^2 - e_2^2$ plane. *Backus and Gilbert* [1970] have termed such graphs *tradeoff curves*. One can easily show the following:

$$\begin{aligned} \frac{\partial}{\partial \theta} [e_1^2] &> 0 \\ \frac{\partial}{\partial \theta} [e_2^2] &< 0 \\ \frac{d[e_2^2]}{d[e_1^2]} &= -\cot \theta \\ \frac{d^2[e_2^2]}{d[e_1^2]^2} &> 0 \quad 0 < \theta < \pi/2 \end{aligned}$$

From these relations we infer that the tradeoff curve between e_1^2 and e_2^2 is monotonically decreasing and convex towards the origin and that θ is the acute angle between the tangent to the tradeoff curve and the e_2^2 axis. These and other qualitative features of the tradeoff curve have been given by *Backus and Gilbert* [1970], who point out that it is best to avoid solutions corresponding to extremal values of θ . Clearly, the generalized inverse solution is a poor choice on the tradeoff curve. The optimal point is where $d[e_2^2]/d[e_1^2] = -1$ corresponding to the solution Eq. (16) given by the stochastic inverse.

APPENDIX

Construction of the operator C_{SS}

In this appendix we present a useful form of the solution auto-correlation operator C_{SS} based on its Karhunen-Loeve expansion Eq. (11). For generality we restrict the definition to an arbitrary semi-open subinterval $(a, b]$ of $[0, 1]$, $0 \leq a < b \leq 1$. The construction yields C_{SS} as a member of a one-parameter family of smoothing operators with norm (and therefore gain) less than or equal to one.

The operator C_{SS} is entirely determined by the specification of an orthonormal basis $\{f_n : n = 1, 2, 3, \dots\}$ and a set of spectral coefficients $\{\alpha_n : n = 1, 2, 3, \dots\}$. Let Λ be the Sturm-Liouville operator: $\Lambda(x) = d/dx [p(x) d/dx] - q(x)$, $x \in (a, b]$. The system with homogeneous boundary conditions

$$\Lambda(x) f(x) + k^2 p(x) f(x) = 0,$$

$$\begin{aligned} \left[\frac{d}{dx} f(x) + \lambda f(x) \right]_{x=a} &= 0 \\ \left[\frac{d}{dx} f(x) + \mu f(x) \right]_{x=b} &= 0 \end{aligned} \quad (A1)$$

generates a set of eigenvectors $\{f_n\}$, taken to be normalized, that is complete on the interval $(a, b]$ [Morse and Feshbach, 1953]. The eigenvalues k_n^2 can be ordered as a continuously increasing sequence; $k_1^2 < k_2^2 < \dots < k_n^2 < \dots$. With this ordering the number of nodes in the eigenfunctions f_n between a and b also forms a continuously increasing sequence [Morse and Feshbach, 1953, p. 722]. This provides the necessary ordering of the basis f_n discussed in the text. For a given scalar value of the parameter k we specify the spectral coefficients by

$$\alpha_n^2 = \frac{k^2}{k^2 + k_n^2} \quad n = 1, 2, 3, \dots \quad (A2)$$

Particularized in this way C_{SS} has the following desired properties:

- (i) $\|C_{SS}\| \leq 1$
- (ii) $\lim_{k \rightarrow \infty} \|(I - C_{SS}) \cdot m\| = 0, n \in M$
- (iii) for k finite, C_{SS} behaves as a low pass filter, i.e., $\alpha_1^2 > \alpha_2^2 > \dots > \alpha_n^2 > \dots$

Then the kernel of C_{SS} satisfies the inhomogeneous system

$$\begin{aligned} \left[1 - \frac{1}{k^2 \rho(x)} \Lambda(x) \right] C_{SS}(x_0, x) &= \frac{\delta(x - x_0)}{\rho(x)} \\ \left[\frac{d}{dx} C_{SS}(x_0, x) + \lambda C_{SS}(x_0, x) \right]_{x=a} &= 0 \\ \left[\frac{d}{dx} C_{SS}(x_0, x) + \mu C_{SS}(x_0, x) \right]_{x=b} &= 0 \end{aligned} \quad (A3)$$

As an example we solve this system for the special case $\rho(x) = \rho(x) = x^2$, $q(x) = 0$. In this case the system Eq. (A1) has a singular point at $x = 0$. Solutions to Eq. (A1) are the spherical Bessel functions of angular order zero. The solution to Eq. (A3) is

$$C_{SS}(x, x_0) = \frac{k}{2xx_0} \left\{ e^{-k|x-x_0|} + \frac{1}{D} \left[A e^{-k(b-a)} \cosh k(x_0 - x) + B \cosh k(a + b - x_0 - x) + C \sinh k(a + b - x_0 - x) \right] \right\} \quad (A4)$$

with

$$A = [1 - a(k + \lambda)] [1 + b(k - \mu)]$$

$$B = \lambda a + \mu b - [k^2 - k(\lambda + \mu) + \lambda \mu] ab - 1$$

$$C = k(b - a)$$

$$D = [1 - \lambda a - \mu b + (\lambda \mu - k^2) ab] \sinh k(b - a) - k[b - a + (\mu - \lambda) ab] \cosh k(b - a)$$

Figure 1 shows the kernel of C_{SS} on the interval $(0, 1]$, centered at $x_0 = 0.5$. Here we take $1/\mu = 1/\lambda = 0$ and display the kernel for k values of 5, 20, and 50.

ACKNOWLEDGMENTS

This research was supported by the Advanced Research Project's agency of the Department of Defense and was monitored by the Air Force Office of Scientific Research under Contract No. F44620-69-C-0067

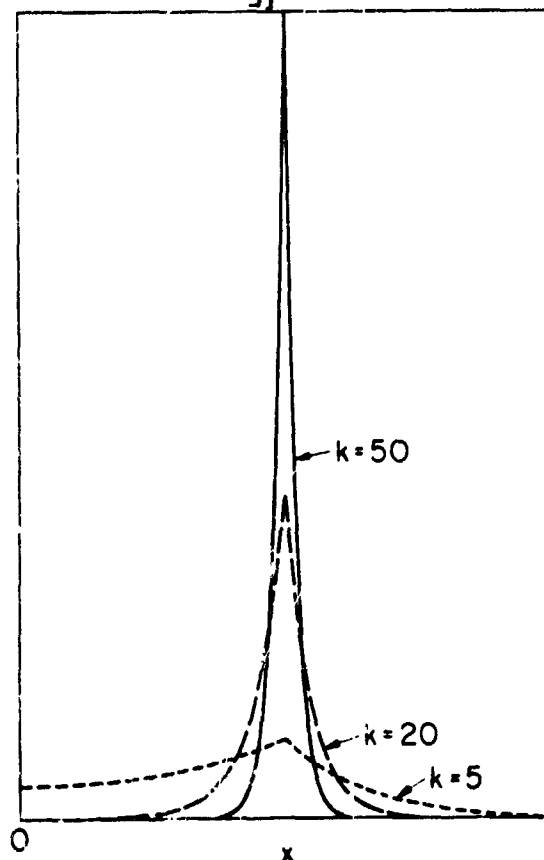


Figure 1.— The integral kernel of C_{SS} centered at $x_0 = 0.5$ for the case $1/\lambda = 1/\mu = 0$ and for k values of 5, 20, and 50.

DISCUSSION

Gross: Apparently the variance of what you want, and the resolution, cannot be independently minimized so that one forms a sum with a weighting function [and for this particular parameter] there is an infinite set of solutions, and one chooses the particular value of interest. What confidence does one have in this solution?

Jordan: This particular solution represents a point on the tradeoff curve. The second minimization shown, where we cast the measure of our error in terms of the sum of two terms, is analogous to the same minimization that Parker and Johnson talked about. The uncertainty in the resolvable part of the solution is given by the second term of this sum. The error due to finite resolution is expressed in the first term. Given only the observations, we can compute the second term, but not the first.

SEISMOLOGY

BIBLIOGRAPHY

- Anderssen, R. C., The Character of Non-uniqueness in the Conductivity Modeling Problem of the Earth, Pure Appl. Geophys., **80**, 238-259, 1970.
- Backus, G. E., Proc. Nat. Academy Sci. (USA), Inference from Inadequate and Inaccurate Data, I, **65**, 1-7, 1970. a.
- Backus, G. E., Inference From Inadequate and Inaccurate Data II, Proc. Nat. Acad. Sci. (USA), **65**, 281-287, 1970. b.
- Backus, G. E., and Gilbert, J. F., Numerical Applications of a Formalism for Geophysical Inverse Problems, Geophys. J. Roy. Astron. Soc., **13**, 247-276, 1967.
- Backus, G. E., and Gilbert, J. F., The Resolving Power of Gross Earth Data, Geophys. J. Roy. Astron. Soc., **16**, 169-205, 1968.
- Backus, G. E. and Gilbert, J. F., Constructing P-Velocity Models to Fit Restricted Sets of Travel-Time Data, Bull. Seismol. Soc. Amer., **59**, 1407-1414, 1969.
- Backus, G. E., and Gilbert, J. F., Uniqueness in the Inversion of Inaccurate Gross Earth Data, Phil. Trans., **266**, 123-192, March 5, 1970.
- Bailey, R. C., Inversion of the Geomagnetic Induction Problem, Proc. Roy. Soc., **315**, 185-194, 1970.
- Banks, J. R., Geomagnetic Variations and the Electrical Conductivity of the Upper Mantle, Geophys. J. Roy. Astron. Soc., **17**, 457-487, 1969.
- Bateman, H., The Solution of the Integral Equation which Connects the Velocity of Propagation of an Earthquake Wave in the Interior of the Earth with the Times which the Disturbance takes to Travel to the Different Stations on the Earth's Surface, Phil. Mag., **19**, 576, 1910.
- Ben-Israel, A. and Charnes, A., Contributions to the Theory of Generalized Inverses, SIAM J. Appl. Math., **11**, 667-699, 1963.
- Bjerhammar, A., Rectangular Reciprocal Matrices with Special Reference to Geodetic Calculations, Bull. Geodesique, 188-220, 1951.
- Bullen K. E., An Introduction to the Theory of Seismology, Cambridge University Press, 1963.
- Cabayan, H. S., and Belford, G. G., On Computing a Stable Least Squares Solution to the Inverse Problem for a Planar Newtonian Potential, SIAM J. Appl. Math., **20**, 51-61, 1971.
- Claerbout, J. F., Synthesis of a Layered Medium from its Acoustic Transmission Response, Geophysics, **33**, 264-269, April 1968.
- Claerbout, J. F., and Landers, T., Effects of Thin Soft Layers on Body Waves, Bull. Seis. Soc. of Amer., **59**, 2071-2078, Oct. 1969.
- Courant, R. and Hilbert, D., Methods of Mathematical Physics, Interscience; New York, 1965.

- Davenport, W. B., and Root, W. L., Random Signals and Noise, McGraw-Hill, New York, 1958.
- Dorman, J., Ewing, M., and Oliver, J., Study of Shear-Velocity Distribution in the Upper Mantle by Mantle Rayleigh Waves, Bull. Seismol. Soc. Amer., 50, 87-115, 1960.
- Dunford, N., and Schwartz, J. T., Linear Operators, (I), Interscience Publishers, New York, 1963.
- Evernden, J. F., and Clark, D. M., Study of Teleseismic P I-Travel Time Data, Phy. Earth Planet. Interiors, 4, 1-23, 1970.
- Franklin, J. N., Well-posed Stochastic Extensions of Ill-Posed Linear Problems, J. Math. Anal. Appl., 31, 682-710, 1970.
- Gelfand, I. M., and Levitan, B. M., On the Determination of a Differential Equation from its Spectral Function, Translation in American Mathematical Society Translations, Series 2, No. 1, 1955.
- Gerver, M. L., and Markushevich, V. M., Determination of the Velocity of Propagation of Seismic Waves from the Travel-Time Curve, (in Russian), Report All USSR, Order #2726, 3-51, 1965.
- Gerver, M. L., and Markushevich, V. M., Determination of a Seismic Wave Velocity from the Travel-Time Curve, Geophys. J. R. Astron. Soc., 11, 165-173, 1966.
- Gerver, M. L., and Markushevich, V. M., On the Characteristic Properties of Travel-Time Curves, Geophys. J. Roy. Astron. Soc., 13, 241-246, 1967.
- Gerver, M. L., and Markushevich, V. M., Vychislitel'naya Seismologiya, III, Trudy Instituta Fiziki Zemli, (in press).
- Gilbert, J. F., Ranking and Winnowing Gross Earth Data for Inversion and Resolution, Geophys. J. Roy. Astron. Soc., 23, 125-128, June 1971.
- Greville, T. N. E., The Pseudo-Inverse of a Rectangular or Singular Matrix and its Applications to the Solution of Systems of Linear Equations, SIAM Rev., 1, 38-43, 1959.
- Herglotz, V. G., Über das Benndorfsche Problem der Fortpflanzungsgeschwindigkeit der Erdbebenstrahlen, Phys. Z., 8, 145-147, 1907.
- Herrin, E., Tucker, W., Taggart, J., Gordon, D. W., and Lobdell, J. L., Estimation of Surface-Focus P Travel Times, Bull. Seismol. Soc. Amer., 58, 1273-1292, 1968.
- Jeffreys, H., The Earth, 4th ed., Cambridge Univ. Press, London and New York, 1962.
- Jordan, T. H., and Franklin, J. N., Optimal Solutions to a Linear Inverse Problem in Geophysics, Proc. Nat. Acad. Sci. USA, 68, 291-293, 1971.
- Jordan, T. H., and Anderson, D. L., The Synthesis of Velocity and Density Models for the Earth, (in preparation).
- Keilis-Borok, V. I., and Yanovskaya, T. B., Inverse Problems of Seismology, Geophys. J. Roy. Astron. Soc., 13, 223-234, 1967.
- Keilis-Borok, V., Profile Inversion Algorithms Used in Seismology, Lecture at MIT Geophysics Dept., Jan. 21, 1969.

Krein, M. G., The Ideas of P. L. Gebysev and A. A. Markov in the Theory of Limiting Values of Integrals and their Further Development, Uspehi mat. nauk, 6, 3-120, 1951; also found in Amer. Math. Soc. Trans., Series 2, 1-121, 1959.

Lancsoz, C., Linear Differential Operators, D. Van Nostrand Co., London, 1961.

Loeve, M., Probability Theory, D. Van Nostrand, Princeton, N. J., p. 478, 1955.

Milder, D. M., Ray and Wave Invariants for SOFAR Channel Propagation, J. Acoust. Soc. Amer., 46, 1259-1263, 1969.

Milder, D. M., and Wells, W. W., Acoustic Holography with Crossed Linear Arrays, presented at Conference on Holography and the Computer, Houston, Texas, Dec. 1969.

Moore, E. H., Bull. Amer. Math. Soc., 25, 394-395, 1920.

Morse, P. M., and Feshbach, H., Methods of Theoretical Physics (I), ch. 6, McGraw-Hill, New York, 1953.

Mostalavi, M., and Mittra, R., Remote Probing of Inhomogeneous Media Using Parameters Optimization Techniques, Scientific Report No. 14, Antenna Lab., U. of Illinois, April 1970.

Parker, R. L., The Inverse Problem of Electrical Conductivity in the Mantle, Geophys. J. Roy. Astron. Soc., 22, 121-138, 1970.

Penrose, R., A Generalized Inverse for Matrices, Proc. Cambridge Phil. Soc., 51, 406-413, 1955.

Phinney, R. A., Reflection of Acoustic Waves from a Continuously Varying Interfacial Region, Rev. Geophys. Space Phys., 8, 517-532, Aug. 1970.

Phinney, R. A., and Anderson, D. L., On the Radio Occultation Method for Studying Planetary Atmospheres, J. Geophys. Res., 73, 1819-1827, 1968.

Press, F., Density Distribution in the Earth, Science, 160, 1218-1221, June 14, 1968.

Press, F., Earth Models Obtained by Monte Carlo Inversion, J. Geophys. Res., 73, 5223-5234, 1968.

Press, F., Earth Models Consistent with Geophysical Data, Phys. Earth Planet. Interiors, 3, 3-22, 1970. a.

Press, F., Regionalized Earth Models, J. Geophys. Res., 75, 6575-6581, 1970. b.

Press, F., and Biehler, S., Inferences on Crustal Structure Velocities and Densities From P Wave Delays and Gravity Anomalies, J. Geophys. Res., 69, 2979-2996, 1964.

Prilepko, A. I., Uniqueness of the Solution of the External Inverse Problem of Newtonian Potentials, Diff. Uravn., 2, 107-124, 1966.

Rapoport, I. M., On the Planar Inverse Problem of Potential Theory, Doklady Akad. Nauk SSSR, 28, 305-310, 1940.

Schaubert, D., and Mittra, R., Remote Probing Methods for the Determination of the Profile of Inhomogeneous Media, Scientific Report No. 15, Antenna Lab., U. of Illinois, May 1970.

Smith, M. L., and Franklin, J. N., A Geophysical Application of Generalized Inverse Theory, J. Geophys. Res., 74, 2783-2785, 1969.

Tseng, Ya. Yu., Generalized Inverses of Unbounded Operators Between Two Unitary Spaces, Doklady Akad. Nauk SSSR, 67, 431-434, 1949.

Twomey, S., On the Numerical Solution of Fredholm Integral Equations of the First Kind by the Inversion of Linear Systems Produced by Quadrature, J. Assoc. Comp. Mach., 10, 79-101, 1963.

Ware, J. A., and Aki K., Continuous and Discrete Inverse-Scattering Problems in a Stratified Elastic Medium. I. Plane Waves at Normal Incidence, J. Acoust. Soc. Amer., 45, 911-921, 1969.

Weidelt, P., Doctoral Dissertation, Göttingen University, 1970.

Wells, W. A., Acoustic Imaging with Linear Transducer Arrays, preprint prepared for the 2nd International Symposium on Acoustical Holography, 1970.

Whittaker, E. T., Analytical Dynamics, 4th ed., Cambridge University Press, 1937.

Wiechert, V. E., and Geiger, L., Bestimmung des Weges der Erdbebenwellen im Erdinnern, Phys. Z., XI, 294-311, 1910.

Wiggins, R. A., Terrestrial Variational Tables for the Periods and Attenuation of the Free Oscillations, Phys. Earth Planet Interiors, 1, 201-266, 1968.

Wiggins, R. A., Monte Carlo Inversion of Body-Wave Observations, J. Geophys. Res., 74, 3171-3181, June 15, 1969.

Wiggins, R. A., Terrestrial Variational Tables for the Periods and Attenuation of the Free Oscillations, Phys. Earth Planet Interiors, 1, 201-266, 1968.

Wiggins, R. A., The General Linear Inverse Problem: Implications of Surface Waves and Free Oscillations on Earth Structures, Rev. of Geophys. and Space Phys., 10, Feb. 1972

Wilkinson, J. H., The Algebraic Eigenvalue Problem, Oxford University Press, 1965.

8. MISCELLANEOUS CONTRIBUTIONS

The papers presented in this chapter are accounts of voluntary contributions presented during the two workshop sessions. They expand on points made in previous chapters or describe some inversion problems not properly falling in one of the major disciplines considered at the Workshop. The last section of the Chapter is an edited account of the more important discussions at the Workshop panel sessions.

J. Shmoys chaired both Workshop sessions. The published papers are authors' summaries of their oral comments or edited transcriptions of their comments.

SOME METHODS FOR DETERMINING THE PROFILE FUNCTIONS OF INHOMOGENEOUS MEDIA

R. Mittra, D. H. Schaubert and M. Mostafavi

University of Illinois

N73-11622

ABSTRACT

The problem of determining the electric permittivity of an inhomogeneous dielectric medium is considered. Two analytical methods and one numerical method for solving this problem are presented. All these methods utilize the reflection coefficient measured as a function of the angular spectrum variable β . This allows one to make measurements at a fixed frequency, and these methods, therefore, can be used when the medium being studied is dispersive. A few examples are provided and some comments are made regarding the stability of solutions obtained using the various methods.

INTRODUCTION

The problems associated with making remote probing measurements are known to workers in many fields. Engineers, physicists, biologists, geologists, ecologists, and many others have been confronted with the problem of determining the nature of an object when it is impossible or undesirable to place measuring devices directly at the point of interest. It then becomes necessary to make measurements at some point removed from the object and to determine the nature of the object from these measurements.

The general remote probing problem is very difficult to solve. This leads most authors to consider a simplified version of the problem, that of determining the characteristics of an inhomogeneous medium that is uniform in two dimensions and varies in the other dimension. Also, the variations are assumed to result from only a few parameters.

The particular type of problem discussed here is that of determining the electrical permittivity of a medium that is uniform in the x and y directions and varies according to the profile function $\kappa(z)$ in the z direction.

Using the techniques of Gelfand and Levitan [1955] and Marchenko [1955], many authors, including Becker and Sharpe [1969] and Kay [1959], have reported the solution of this type of problem. All of their work, however, requires that the response of the medium be known as a function of the temporal frequency. It is also required that the medium itself be independent of frequency—that is, nondispersive. Often, however, it is desirable to make measurements at a single frequency. Therefore, in this paper we consider the remote probing problem in the angular spectrum domain. Here the independent variable β is the spatial frequency, which can be associated with $k \sin \theta$ where k is the free space wave number and θ is the angle of incidence of a probing plane wave.

The technique by which the response of the medium is measured is to illuminate the medium using a spatially confined source. The total field at the interface is then measured. The information is then converted to the angular spectrum domain by taking the Fourier transform with respect to the x direction.

In the next section, the differential equation describing the problem is derived. Three different methods of obtaining the profile are then discussed. The first method makes use of the form of the equation in the limit of low temporal frequencies [Schaubert and Mittra, 1970]. The second method can be used for arbitrary frequency and is based on a new representation for the wave function which is appropriate for the angular spectrum domain. The third method is a direct numerical approach which uses parameter optimization techniques [Mostafavi and Mittra, 1970].

FORMULATION OF THE PROBLEM

The geometry of the problem to be considered is shown in figure 1. Note that there are no variations with respect to the y direction. The dielectric is also uniform in the x direction and varies in the z direction as $\kappa(z)\epsilon_0$ where ϵ_0 is the permittivity of free space. The permeability is assumed to be constant and equal to that of free space. It is assumed that a known electromagnetic wave of a single frequency is incident from the left on the air-dielectric interface at $z = 0$ and that the tangential fields at the surface of the dielectric can be measured for all values of x . Of course, the actual measurement could be performed at some plane $z = \alpha$, $\alpha < 0$, and then the fields at the interface could be computed since the propagation in free space is completely known.

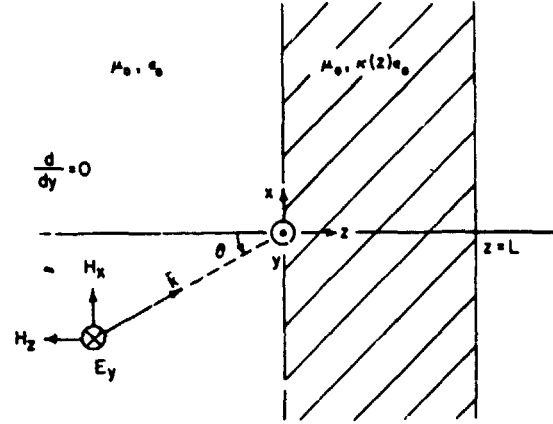


Figure 1. - Remote probing problem in the angular spectrum domain.

If the incident field is polarized such that the components E_x , E_z , and H_y are all zero, then propagation in the medium is completely described by the scalar equation

$$\nabla^2 v(x, z) + \omega^2 \mu_0 \kappa(z) \epsilon_0 v(x, z) = 0 \quad (1)$$

(ω is the angular frequency of the incident wave). The function $v(x, z)$ may be identified with E_y or H_z , whichever is more convenient for the problem at hand.

To obtain the differential equation in the angular spectrum domain, we introduce the Fourier transform relations:

$$u(\beta, z) = \int_{-\infty}^{\infty} v(x, z) \exp(-j\beta x) dx \quad (2)$$

$$v(x, z) = \frac{1}{2\pi} \int_{-\infty}^{\infty} u(\beta, z) \exp(j\beta x) d\beta$$

Then Eq. (1) becomes

$$\frac{d^2 u(\beta, z)}{dz^2} + \left[\omega^2 \mu_0 \kappa(z) \epsilon_0 - \beta^2 \right] u(\beta, z) = 0 \quad (4)$$

This is the differential equation we are seeking. Recall that ω , μ_0 , and ϵ_0 are constants, β is the independent variable, and $\kappa(z)$ is the unknown profile function.

METHOD 1: A LINEAR INTEGRAL EQUATION

In this method, we assume that the medium is terminated at $z = L$ by a perfect electric conducting wall (see fig. 2). This simplifies the expressions by causing some of the boundary terms to vanish.

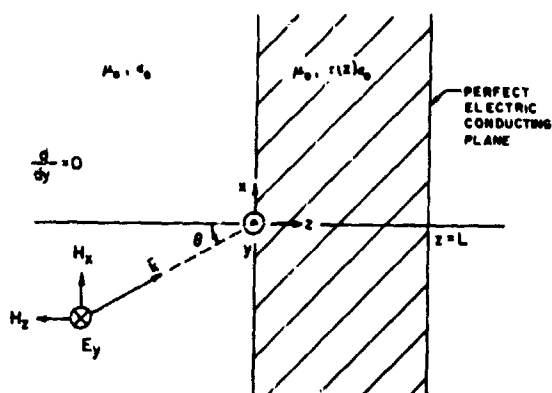


Figure 2. - Geometry of the remote probing problem with a perfectly conducting plane at $z = L$.

We look now at Eq. (4) and consider the special case of $\omega = 0$.

$$\frac{d^2 S(\beta, z)}{dz^2} - \beta^2 S(\beta, z) = 0 \quad (5)$$

Note that $S(\beta, z)$ is the static solution for $H_z(\beta, z)$ and is independent of $\kappa(z)$. Therefore $S(\beta, z)$ can be calculated.

$$S(\beta, z) = S_0 \sinh [\beta(z - L)] \quad (6)$$

We can now use Eqs. (4), (5), and (6) and the fact that both $u(\beta, z)$ and $S(\beta, z)$ vanish at $z = L$, to arrive at the following integral equation for $\kappa(z)$.

$$\omega^2 \mu_0 \epsilon_0 \int_0^L \kappa(z) u(\beta, z) \sinh[\beta(z - L)] dz = -\beta \cosh[\beta L] u(\beta, 0) - \sinh[\beta L] \frac{du(\beta, 0)}{dz} \quad (7)$$

Two points should be noted about Eq. (7). First, the right-hand side is related to the tangential fields at the surface of the dielectric and, thus, can be determined from the input impedance. Second, Eq. (6) is nonlinear because $u(\beta, z)$ depends on $\kappa(z)$. Therefore, there is no systematic procedure that can be used to find $\kappa(z)$ from Eq. (7). To circumvent this problem, a linear form of the integral equation is obtained by expanding $u(\beta, z)$ in a power series of ω , the angular frequency.

$$u(\beta, z) = \sum_{n=0}^{\infty} C_n(\beta, z) \omega^n \quad C_n(\beta, z) = \frac{1}{n!} \frac{\partial^n u(\beta, z)}{\partial \omega^n} \bigg|_{\omega=0} \quad (8)$$

It can be proven that for ω sufficiently small, no surface wave of this polarization will propagate and that Eq. (8) is a uniformly convergent series. This allows us to substitute Eq. (8) into (7) and interchange the order of summation and integration. Equating the coefficients of like powers of ω yields a series of equations. The ω^2 term, in particular, yields

$$\mu_0 \epsilon_0 \int_0^L \kappa(z) C_0(\beta, z) \sinh[\beta(z - L)] dz = -\beta \cosh[\beta L] C_2(\beta, 0) - \sinh[\beta L] \frac{dC_2(\beta, z)}{dz} \bigg|_{z=0} \quad (9)$$

To simplify Eq. (9), note that $C_0(\beta, z) = u(\beta, z) \big|_{\omega=0}$. But, for $\omega = 0$, Eq. (4) becomes identical with Eq. (5), and $u(\beta, z)$ becomes identical with $S(\beta, z)$. Therefore, Eq. (9) can be written

$$\mu_0 \epsilon_0 \int_0^L \kappa(z) \sinh^2 \beta(z-L) dz = -\beta \cosh[\beta L] C_2(\beta, 0) - \sinh[\beta L] \left. \frac{dC_2(\beta, z)}{dz} \right|_{z=0} \quad (10)$$

This is the linear integral equation that $\kappa(z)$ must satisfy.

It should be noted that Eq. (10) is exact. No approximations were made in the expansion of $u(\beta, z)$ or in the matching of coefficients. It should also be noted that the linear equation is not obtained without cost. The right-hand side of Eq. (10) involves the second derivative with respect to frequency of the tangential fields. This requires that the probing be done using frequencies in the neighborhood of zero. This, however, may not be a limitation since some remote probing, such as the determination of the electrical properties of the earth, is often carried out at very low frequencies.

Finally, it should be noted that the satisfaction of Eq. (10) is sufficient to guarantee that the equations obtained by matching other powers of ω are all satisfied.

One of the greatest advantages of the linear form of the integral equation is that the solution can be shown to be unique. To do so, we must show that the homogeneous equation

$$\int_0^L f(z) \sinh^2 [\beta(z-L)] dz = 0 \quad (11)$$

has only the trivial solution. Rewriting Eq. (11), we have

$$\frac{1}{2} \int_0^L f(z) \exp[2\beta(z-L)] dz + \frac{1}{2} \int_0^L f(z) \exp[-2\beta(z-L)] dz - \int_0^L f(z) dz = 0 \quad (12)$$

which says that a linear combination of three terms must vanish for all values of β . Clearly the last term in Eq. (12) is independent of β ; considering its first term, if $|f(z)| \leq M$ for $0 \leq z \leq L$,

$$\left| \int_0^L f(z) \exp[2\beta(z-L)] dz \right| \leq M \frac{1 - \exp[-2\beta L]}{2\beta} \quad (13)$$

Therefore, for large β , the first integral is of the order of β^{-1} . Now consider the second term in Eq. (12).

$$\int_0^L f(z) \exp[-2\beta(z-L)] dz = \exp[2\beta L] \int_0^L f(z) \exp[-2\beta z] dz \quad (14)$$

The integral on the right-hand side is the finite Laplace transform integral. Hence, if $f(z)$ is a Laplace transformable function, the second term of Eq. (12) behaves as $\exp[2\beta L]$ for large β , since the integral $\int_0^L f(z) \exp(-2\beta z) dz$

is algebraic for large β . Therefore, Eq. (12) can be satisfied for all β only if $f(z)$ is identically zero. Hence, the solution of Eq. (10) is unique within the class of bounded, Laplace transformable functions.

A final comment about the numerical behavior of the solution of Eq. (10) is needed. The matrices obtained by various quadratures on Eq. (10) were found to be ill-conditioned. That is, small deviations in the data appearing on the right-hand side of the equation lead to large deviations in the calculated profile function. This limits the application of this method in practical problems. However, Eq. (10) was proven to be nonsingular; hence, better inversion methods than were used in our investigation could lead to usable results.

METHOD 2: A NEW REPRESENTATION

Recall that method 1 required the angular spectrum response of the medium to be obtained for angular frequencies in the vicinity of $\omega = 0$. The method presented in this section can be used for any fixed nonzero value of ω .

Before continuing, we introduce an equivalent form of Eq. (4), which will be used as the basic differential equation in this section

$$u''(\gamma, z) + \gamma^2 u(\gamma, z) = Q(z) u(\gamma, z) \quad (15)$$

where

$$\begin{aligned} \gamma^2 &= k^2 - \beta^2 \\ Q(z) &= -k^2 [\kappa(z) - 1] \\ k^2 &= \omega^2 \mu_0 \epsilon_0 \end{aligned}$$

The prime denotes differentiation with respect to z . Note that if we determine $Q(z)$, $\kappa(z)$ can easily be found.

We will assume that $Q(z)$ is bounded and that

$$\int_0^\infty z |Q(z)| dz < \infty \quad (16)$$

$$\lim_{z \rightarrow \infty} u(\gamma, z) = \exp[i\gamma z] \quad (17)$$

We then write the solution to Eq. (15) in the form

$$u(\gamma, z) = \exp[i\gamma z] + \frac{1}{2\pi i} \int_{-\infty}^{\infty} u(\eta, z) \exp[i(\gamma + \eta)z] (\gamma + \eta)^{-1} P(\eta) d\eta \quad (18)$$

The associated profile function is then given by the expression

$$Q(z) = \frac{1}{\pi i} \int_{-\infty}^{\infty} [iu'(\eta, z) - \eta u(\eta, z)] \exp[i\eta z] P(\eta) d\eta \quad (19)$$

The function $P(\eta)$ in Eqs. (18) and (19) contains information about the medium being studied. We can determine $P(\eta)$ from a knowledge of $u(\gamma, 0)$ by considering Eq. (18) for $z = 0$. That is, the equation

$$u(\gamma, 0) = 1 + \frac{1}{2\pi i} \int_{-\infty}^{\infty} u(\eta, 0) (\gamma + \eta)^{-1} P(\eta) d\eta \quad (20)$$

is a linear integral equation for $P(\eta)$.

Using these equations, we obtain the unknown functions $Q(z)$ by the following steps:

1. Knowing $u(\gamma, 0)$, solve Eq. (20) for $P(\eta)$.
2. For various fixed values of z , solve Eq. (18) for $u(\gamma, z)$.
3. Obtain $Q(z)$ from Eq. (19).

It can be shown that Eq. (18) is a solution to (15). This is done by substituting Eq. (18) into (15), yielding an integral equation of the form

$$T(\gamma, z) = \frac{i}{2\pi i} \int_{-\infty}^{\infty} T(\eta, z) \exp[i(\gamma + \eta)z] (\gamma + \eta)^{-1} P(\eta) d\eta \quad (21)$$

where $T(\gamma, z) = u''(\gamma, z) + \gamma^2 u(\gamma, z) - Q(z) u(\gamma, z)$. For functions of the type encountered in this problem, Eq. (21) has only the solution $T(\gamma, z) \equiv 0$. Therefore, Eq. (15) is satisfied.

Using an example due to Sharpe [1963], the above method was tested for numerical stability. In this case, the field at the interface was given by

$$u(\gamma, 0) = \frac{(\gamma + i)^2}{[\gamma + (\sqrt{7}/2) + (i/2)] [\gamma - (\sqrt{7}/2) + (i/2)]} \quad (21)$$

Figure 3 indicates the close comparison between the calculated result and the analytical result obtained by Sharpe.

METHOD 3: PARAMETER OPTIMIZATION

In contrast to the two methods presented above, the third method is numerical in nature. This method is more general than the previous ones, but it does lack the analytical niceties that are present in those methods. For instance,

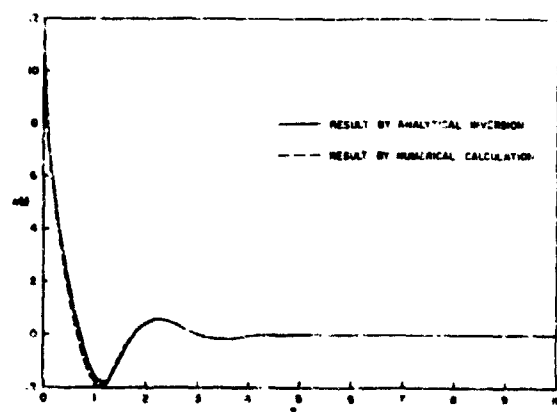


Figure 3. - Results obtained when method 2 was applied to Sharpe's example.

it is impossible to prove uniqueness of the solution. Nevertheless, this method can use data obtained in the angular spectrum domain or the frequency domain, or both, and it can be generalized to a wider class of geometries—for example, circular columns and sheath configurations in the neighborhood of space vehicles. In addition, this method is numerically stable.

The method is based on the parameter optimization approach commonly used in systems theory. The method seeks to minimize a performance index function F , which is defined as the L_2 norm of the error of the response.

$$F(\kappa_1, \kappa_2, \dots, \kappa_m) = \sum_{i=1}^P \sum_{j=1}^Q |u^g(\omega_i, \beta_j, \kappa) - u(\omega_i, \beta_j, \kappa_1, \kappa_2, \dots, \kappa_m)|^2 \quad (23)$$

where

κ_i are unknown parameters introduced to describe the medium,

$u^g(\omega_i, \beta_j, \kappa)$ = surface field for the true medium described by a profile function $\kappa(z)$ when measured for angular frequency ω_i and spatial frequency β_j .

$u(\omega_i, \beta_j, \kappa_1, \kappa_2, \dots, \kappa_m)$ = surface field for a trial medium described by the parameters $\kappa_1, \kappa_2, \dots, \kappa_m$ when measured for angular frequency ω_i and spatial frequency β_j .

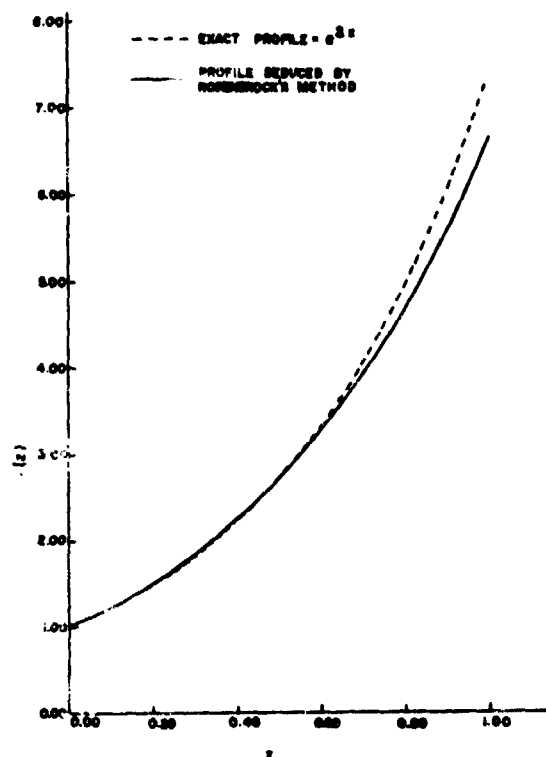


Figure 4.— Profiles of electric constant $\kappa(z) = e^{2z}$ for a slab terminated by a perfect conductor at $z = 1$. Rosenbrock's method.

The terms $u^g(\omega_i, \beta_j, \kappa)$ are the measured values of the field for the medium being investigated.

The strategy of the method is as follows. One begins with a continuous or layered model for the medium involving the parameters $\kappa_1, \kappa_2, \dots, \kappa_m$. With the parameters set equal to some reasonable values, the response of this trial medium for each ω_i and β_j is calculated using Eq. (4) and some standard numerical technique for solving differential equations. This response is the $u(\omega_i, \beta_j, \kappa_1, \kappa_2, \dots, \kappa_m)$ term in Eq. (23). Once the performance index F is calculated, the optimization algorithm determines a new set of parameter values and the process is repeated until a preset stopping criterion is satisfied—for example, F is less than a given error tolerance ϵ .

The optimization technique has been applied to a variety of inversion problems involving inhomogeneous media with continuous as well as discrete variations in the profile function, for both planar and circular geometries.

The numerical calculations were done using two different optimization algorithms. Figure 4 shows the results using Rosenbrock's [1960] method for a profile function $\kappa(z) = \exp(2z)$. Figure 5 shows the results obtained using the conjugate gradients method [Fletcher and Reeves, 1964] on the same profile. Figure 6 shows the results

obtained using the conjugate gradients method when the data was polluted by noise. The very good agreement here indicates that this method is stable enough to be used with laboratory data.

It can be seen that the computed results agree with the true profile better in the vicinity of $z = 0$. This can be explained by the fact the signal returned from the deeper inside of the medium (near the perfect conducting wall) is masked by traveling through the front portion of the medium. Thus, variations occurring deep in the medium have little effect on the fields measured at the interface. Figures 7 and 8 show the results obtained for a profile illuminated from the left and from the right, respectively. In both cases the results are better near the interface than near the perfect conductor.

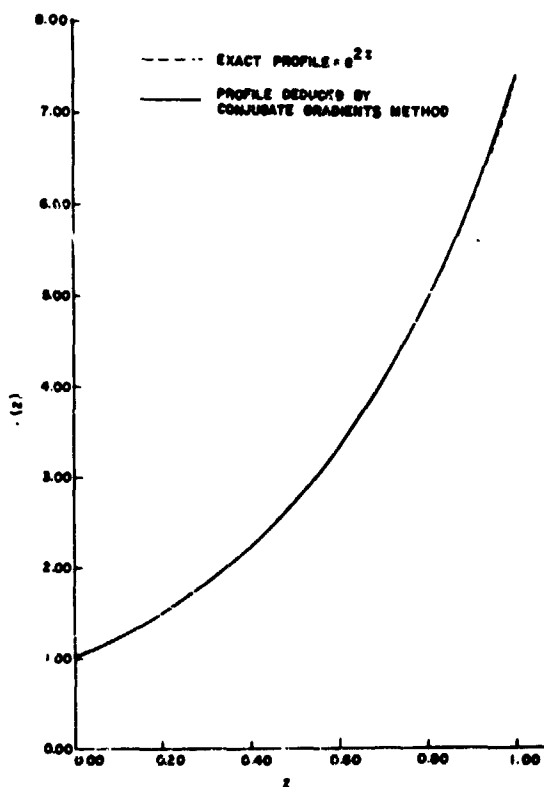


Figure 5.— Profile of dielectric constant $\kappa(z) = e^{2z}$ for a slab terminated by a perfect conductor at $z = 1$. Conjugate gradients method

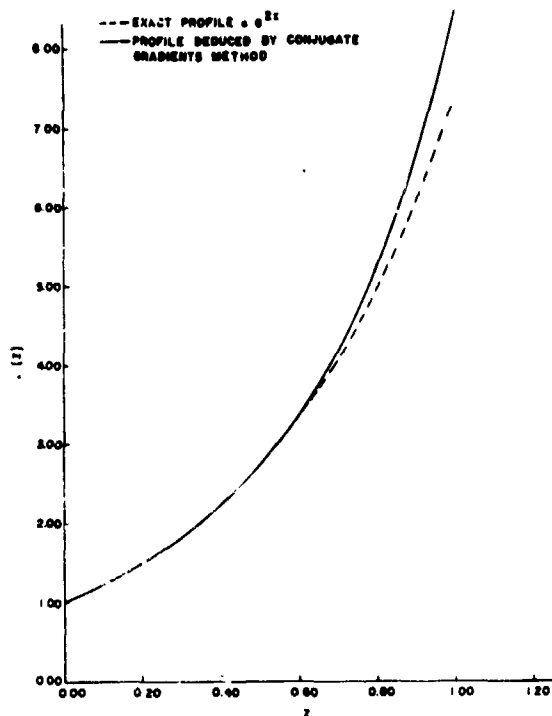


Figure 6.— Profile of dielectric constant $\kappa(z) = e^{2z}$ for a slab terminated by a perfect conductor at $z = 1$. Random noise was added to the simulated data.

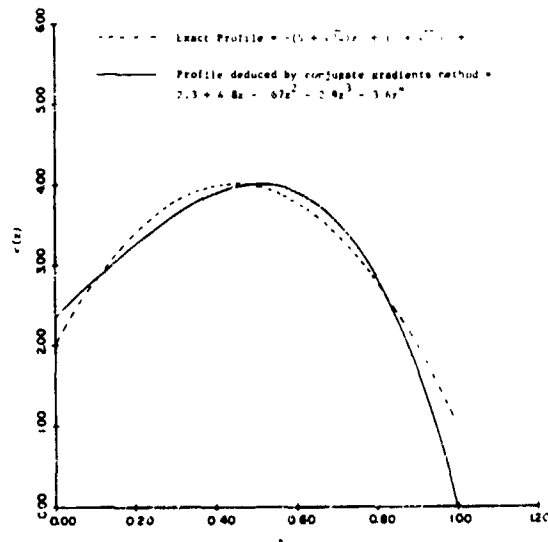


Figure 7.— Profile of dielectric constant $\kappa(z) = -(5 + \sqrt{24})z^2 + (4 + \sqrt{24})z + 2$ for a slab terminated by a perfect conductor at $z = 1$. Plane wave was incident from $-z$ to $+z$ direction. Conjugate gradients method, $k_0 = 1.0$.

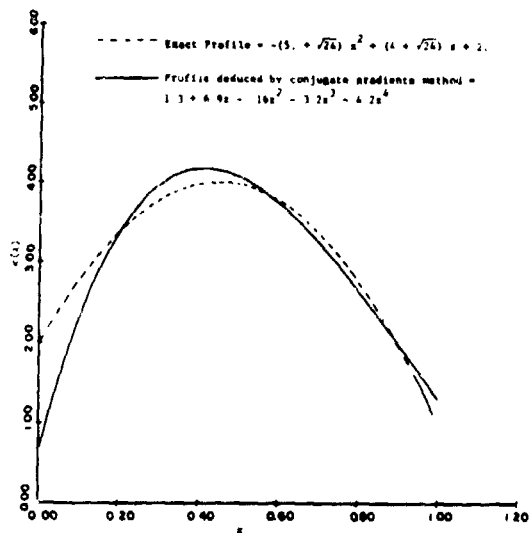


Figure 8.— Profile of dielectric constant $\kappa(z) = -(5 + \sqrt{24})z^2 + (4 + \sqrt{24})z + 2$ for a slab with a perfect conductor at $z = 0$. Plane wave was incident from $+z$ to $-z$ direction. Conjugate gradients method, $k_0 = 1$.

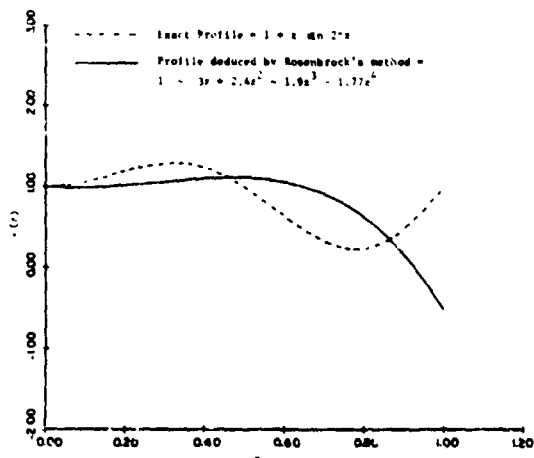


Figure 9.— Profile of dielectric constant $\kappa(z) = 1 + z \sin 2\pi z$ for a dielectric slab. Rosenbrock's method, $k_0 = 1$.

Figure 9 is another example in which variations of the profile function occurring deep in the medium have little effect on the surface fields. This example illustrates the uncertainty that must be attached to the proposed solutions. Although the results are not particularly good, the performance index for this case was very small, less than 10^{-3} , and the amplitudes of the scattered fields agreed to two significant digits.

CONCLUSIONS

The three methods discussed for solving the remote probing problem were developed especially to utilize data obtained in the angular spectrum domain. The first method is a novel technique for extracting an exact linear integral equation from a nonlinear integral equation. The second method, based on a new representation of the solution to the differential equation describing the problem, is more stable numerically than the first.

The third method was seen to be numerical in nature and so lacks many of the analytical niceties that the first two methods possess. However, this method can use data from either the frequency domain or the angular spectrum domain, and it can be used for geometries other than planar. Furthermore, this method is numerically stable.

DISCUSSION

Grossi: Would this be applicable to the case of vertical incidence ionograms, the valley E and F layer?

Mitra: Yes With certain provisions.

Shellman: In the last example, what was the allowable class of functions?

Mitra: We have not found the restrictions yet that one has to put

REFERENCES

1. Gelfand, I. M.; and Levitan, B. M.: On the Determination of a Differential Equation from its Spectral Function. *Amer. Math. Soc. Translations, Series 2*, vol. 1, 1955, pp. 253-304.
2. Marchenko, V. A.: Reconstruction of the Potential Energy from the Phase of Scattered Waves. *Dokl. Akad. Nauk SSSR*, vol 164, pp. 695-698.
3. Becher, W. D.; and Sharpe, C. B.: A Synthesis Approach to Magnetotelluric Exploration. *Radio Sci.*, vol. 4, 1969, pp. 1089-1094.
4. Kay, L.: *On the Determination of the Free Electron Distribution of an Ionized Gas*, New York Univ. Res. Rep. EM-141, Sep. 1959.
5. Schaubert, D. H.; and Mittra, R.: *Remote Probing Methods for the Determination of the Profile of Inhomogeneous Media*, Univ. of Illinois, Antenna Lab., Scientific Rep. no. 15, May 1970.
6. Mostafavi, M.; and Mittra, R.: *Remote Probing of Inhomogeneous Media Using Parameter Optimization Techniques*, Univ. of Illinois, Antenna Lab., Scientific Rep. no. 14, April 1970.
7. Sharpe, C. B.: The Synthesis of Infinite Lines. *Quart. Appl. Math.*, vol. 21, 1963, p. 105.
8. Rosenbrock, H. H.: An Automatic Method for Finding the Greatest or Least Value of a Function. *Computer J.*, Oct. 1960.
9. Fletcher, R.; and Reeves, C. M.: Function Minimization by Conjugate Gradients. *Computer J.*, vol. 7, no. 2, 1964.

k-SPACE FORMULATION OF THE n -DIMENSIONAL SCATTERING PROBLEM

Norbert N. Bojarski[†]

ABSTRACT

N73-11623

The n -dimensional scattering problem is solved by means of a k -space formulation of the field equations, thereby replacing the conventional integral equation formulation by a set of two algebraic equations in two unknowns in two spaces (the constitutive equation being an algebraic equation in x -space). These equations are solved by an iterative method with the aid of the fast Fourier transform (FFT) algorithm connecting the two spaces, requiring very simple initial approximations. Since algebraic and FFT equations are used, the number of arithmetic multiple-add operations and storage allocations required for a numerical solution are reduced from the order of N^2 (for solving the matrix equations resulting from the conventional integral equations) to the order of $N \log_2 N$ and N , respectively (where N is the number of data points required for the specification of the problem). The advantage gained in speed and storage is thus of the order of $N/\log_2 N$ and N , respectively. This method is thus considerably more efficient than the conventional matrix method, and permits exact numerical solutions for much larger problems. Arguments are presented towards the view that the field equations are more fundamental in k -space. The details and some numerical results of the application of this method to the three-dimensional electromagnetic scattering problem are presented as an example.

INTRODUCTION AND BACKGROUND

Consider the n -dimensional scalar, vector, or tensor field $f(\vec{x})$ and the source density $g(\vec{x})$, governed by the linear m th order differential field equation

$$L_f f(\vec{x}) = -L_g g(\vec{x}) \quad (1)$$

where the n -dimensional linear differential scalar, vector, or tensor operators L_f and L_g are of the form

$$L_f = \sum_{\mu=0}^m a_{\mu} \left(\frac{\partial}{\partial x_i} \right)^{\mu} \quad i = 1, 2, \dots, n \quad (2)$$

$$L_g = \sum_{\mu=0}^m b_{\mu} \left(\frac{\partial}{\partial x_i} \right)^{\mu} \quad (3)$$

[†] Physicist; Research Contractor and Consultant to the *Department of Defense*.
16 Circle Drive, Moorestown, New Jersey 08057; Phone: (609) 235-3031

subject to the n -dimensional scalar, vector, or tensor constitutive equation

$$g(\bar{x}) = h(\bar{x})f(\bar{x}) \quad (4)$$

The conventional n -dimensional integral representation of the generalized initial value or (scattering) problem associated with the field Eq. (1) is

$$f(\bar{x}) = \int_D G(\bar{x} | \bar{x}') g(\bar{x}') d^n x' + f^i(\bar{x}) \quad (5)$$

subject to the constitutive Eq. (4); where $f^i(\bar{x})$ is the externally imposed field [initial value, source distribution external to the problem defined by the constitutive Eq. (4)]; D is the domain of nonvanishing $h(\bar{x})$; and $G(\bar{x})$ is the appropriate Green's function, vector, or tensor, satisfying the differential equation

$$L_f G(\bar{x}) = -L_g \delta(\bar{x}) \quad (6)$$

The conventional numerical method of solution of this initial value problem is by means of numerical matrix inversion methods *Harrington* [1968], applied to the Fredholm integral equation of the second kind, formed by combining Eqs. (4) and (5); that is,

$$f(\bar{x}) - \int_D K(\bar{x} | \bar{x}') f(\bar{x}') d^n x' = f^i(\bar{x}) \quad (7)$$

where the integral transform kernel $K(\bar{x} | \bar{x}')$ is given by

$$K(\bar{x} | \bar{x}') \equiv G(\bar{x} | \bar{x}') h(\bar{x}') \quad (8)$$

Such matrix inversion methods require the order of N^2 computer storage allocations, and of the order of from N^2 to N^3 arithmetic multiply-add operations (for conventional matrix inversions and matrix inversions by iteration methods, respectively), for the execution of a numerical solution; N is the number of data points required for the numerical specification of the constitutive Eq. (4), the specification of the nonvanishing portion of $h(\bar{x})$. The practical size limit with state-of-the-art computers is for N of the order of several hundred.

THE k -SPACE FORMULATION OF THE GENERAL INITIAL-VALUE PROBLEM

The k -space representation and solution of the generalized n -dimensional initial-value problem is discussed. The n -dimensional Fourier Transform of the differential field Eq. (1) yields the algebraic scalar, vector, or tensor k -space equation

$$L_f(\bar{k}) F(\bar{k}) = -L_g(\bar{k}) G(\bar{k}) \quad (9)$$

where

$$F(\bar{k}) = \int_{-\infty}^{\infty} e^{i\bar{k} \cdot \bar{x}} f(\bar{x}) d^n x \quad (10)$$

and where, by virtue of Eqs. (2) and (3), the quantities $L_f(\bar{k})$ and $L_g(\bar{k})$ are polynomials of the form

$$L_f(\bar{k}) = \sum_{\mu=0}^m a_{\mu} (i\bar{k})^{\mu} \quad (11)$$

$$L_g(\bar{k}) = \sum_{\mu=0}^m b_{\mu} (i\bar{k})^{\mu} \quad (12)$$

The k -space representation of the generalized n -dimensional initial value problem, consistent with the x -space integral representation in Eq. (5), thus is the algebraic scalar, vector, or tensor equation (vis-a-vis the conventional integral or differential equation)

$$F(\bar{k}) = \Gamma(\bar{k}) G(\bar{k}) + F^i(\bar{k}) \quad (13)$$

subject to the algebraic x -space constitutive Eq. (4); that is,

$$g(\bar{x}) = h(\bar{x}) f(\bar{x}) \quad (14)$$

where

$$\Gamma(\bar{k}) \equiv \frac{L_g(\bar{k})}{L_f(\bar{k})} \quad (15)$$

which clearly can be taken as the Green's function (or vector or tensor) in k -space, that is,

$$\Gamma(\bar{k}) \leftrightarrow -G(\bar{x}) \quad (16)$$

The generalized n -dimensional initial value problem is thus reduced to a set of two algebraic (scalar, vector, or tensor) equations in two unknowns in two spaces — that is, Eqs. (13) and (14).

The unique existence [Papoulis, 1962] of this k -space representation is restricted to media for which

$$\int |h(\bar{x})| d^n x < \infty \quad (17)$$

If $h(\bar{x})$ is in general nonvanishing only in a finite n -dimensional x -domain, then the pair of algebraic equations (13) and (14) can be solved numerically with the aid of the n -dimensional fast Fourier transform (FFT) algorithm [Cooley *et al.*, 1967] as the connection between the two spaces, by the following iterative relaxation method [Varga, 1962], the recursion relationship for which is

$$\begin{aligned} G_n(\bar{k}) &= F(\bar{k} | \bar{x}) g_n(\bar{x}) \\ F_n(\bar{k}) &= \Gamma(\bar{k}) G_n(\bar{k}) + F^i(\bar{k}) \\ f_n(\bar{x}) &= F(\bar{x} | \bar{k}) F_n(\bar{k}) \\ g_{n+1}(\bar{x}) &= \alpha h(\bar{x}) f_n(\bar{x}) + (1 - \alpha) g_n(\bar{x}) \end{aligned} \quad (18)$$

where α is an appropriately chosen relaxation coefficient (best numerical results to date were obtained for $\alpha = 1/2$), and where $F(\bar{k} | \bar{x})$ and $F(\bar{x} | \bar{k})$ designate the fast Fourier transform algorithm operator and its inverse, respectively.

The initial approximation $g_0(\bar{x})$ can be taken as any known simply programmable approximation to the problem, including the trivial case $g_0(\bar{x}) = 0$.

To avoid the numerical difficulties arising from the fictitious periodic nature of the FFT, which is a discrete and finite Fourier transform, vis-à-vis the continuous and infinite Fourier transform implied by Eq. (10), and the possible singularities in the fields, sources, and Green's function (or vector or tensor) in k -space, it becomes necessary to choose an n -dimensional hyper-rectangular box of twice the size (in each dimension) of the smallest hyper-rectangular box in which the nonvanishing $h(\bar{x})$ is imbeddable as the x domain for the FFT, and take the Green's function, vector, or tensor as

$$\Gamma(\mu, \nu, \dots, \sigma) = -\Delta^n x \sum_{\alpha=-\frac{N_1}{2}}^{\frac{N_1}{2}-1} \sum_{\beta=-\frac{N_2}{2}}^{\frac{N_2}{2}-1} \dots \sum_{\gamma=-\frac{N_n}{2}}^{\frac{N_n}{2}-1} w_{N_1}^{-\mu\alpha} w_{N_2}^{-\nu\beta} \dots w_{N_n}^{-\sigma\gamma} G(\alpha, \beta, \dots, \gamma) \quad (19)$$

in the conventional FFT notation [Cooley *et al.*, 1967] where

$$w_N = e^{2\pi i/N} \quad (20)$$

$$\Delta^n x = \Delta x_1 \Delta x_2 \dots \Delta x_n \quad (21)$$

$$N = N_1 N_2 \dots N_n \quad (22)$$

and where appropriate use is made of the periodicity properties of the FFT [Cooley *et al.*, 1967] for both Eq. (19) and the desired placement of $h(\bar{x})$ and $f^I(\bar{x})$ in the hyper-rectangular FFT x domain.

The numerical difficulty arising from the possible singularity of the Green's function at the origin of the x space (that is, $G(\bar{0}) = \infty$) can be alleviated by taking advantage of the appropriate principal value integral representation of the field Eq. (5), such as, [Van Bladel, 1961]

$$f(\bar{x}) = \ell g(\bar{x}) + \rho \int G(\bar{x} | \bar{x}') g(\bar{x}') d^n x' \quad (23)$$

It thus immediately follows that $G(\bar{0})$ for Eq. (19) can be taken as

$$G(\bar{0}) = \frac{\ell}{\Delta^n x} \quad (24)$$

Since algebraic and FFT equations are used, the number of arithmetic multiply-add operations required [Singleton, 1969] for a solution is reduced from the order of N^2 (required for solving the matrix equations resulting from the conventional integral equation representation of the problem) to the order of $N \log_2 N$, and the storage requirement is reduced also from the order of N^2 (required for storing the matrix associated with the matrix method of solution) to the order of N (required for storing the k - and x -space vectors). The advantage gained in speed and storage is thus of the order of $N/\log_2 N$ and N , respectively. This method is thus considerably more efficient, and permits exact numerical solutions for much larger problems. To date, problems of the order of $N = 10^4$ have been successfully solved; and problems of the order of 10^6 seem feasible with state-of-the-art computers [Singleton, 1970].

THE k -SPACE FORMULATION OF THE GENERAL WAVE SCATTERING PROBLEM

For the general n -dimensional wave scattering problem, the (range- and phase-normalized) scattered far fields in x space are in general simply and algebraically related to the n -dimensional Fourier transform of the source distribution $g(\bar{x})$ — that is, $G(\bar{k})$ — induced by the incident field $f^I(\bar{x})$; $G(\bar{k})$ clearly is yielded directly by the iterative solution in Eq. (18) without additional computations. Since this is not the case with the conventional matrix method of solution of the integral equation representation of the scattering problem, the k -space method of solution presented is particularly and additionally attractive when applied to scattering problems.

For the special case of the n -dimensional Helmholtz (time-reduced) wave equations for which

$$L_f = \sum_{i=1}^n \left(\frac{\partial}{\partial x_i} \right)^2 + k_0^2 \quad (25)$$

where $k_0 \equiv \omega/c$, and c is the wave velocity in free space (this deviation from conventional notation is for the purpose of distinction from k , the Fourier transform variable of x), the n -dimensional k -space Green's function, in the notation of (9), (15), and (16), is clearly

$$\Gamma(\bar{k}) = \Phi(\bar{k}) L_g(\bar{k}) \quad \Phi(\bar{k}) \equiv \frac{1}{k^2 - k_0^2} \quad (26)$$

The form of the Green's function $\Phi(\bar{k})$ in k space is clearly invariant to the dimensionality n of the space, which is not the case in x space.

Thus, by virtue of the previously stated relationship between the (range- and phase-normalized) scattered far fields in x space and the source distribution $G(\bar{k})$ in k space, and conservation-of-energy considerations for passive media of finite spatial extent [see the inequality (17)], it follows that the radiation condition for the Helmholtz equation in k space can be stated as

$$\left| \frac{F(\bar{k}_s)}{\Phi(\bar{k}_s)} \right| < \infty \quad |\bar{k}_s| = k_0 \quad (27)$$

where \bar{k}_s is the propagation wave number vector of the scattered far fields; see also (37) through (40).

It can thus be argued that the field equations of mathematical physics are more fundamental in k space, because of the simple algebraic nature of these equations in k space, and the invariance of the form of the k space Green's function of the Helmholtz equation with respect to the dimensionality of the space, particularly when bearing in mind that the Fourier transform is the only transform known for which a fast algorithm exists. (This apparent more fundamental nature is most obvious for quantum mechanics.)

ELECTROMAGNETIC SCATTERING

Three-dimensional electromagnetic monochromatic scattering by passive inhomogeneous media, including perfect conductors, of finite spatial extent and arbitrary shape, is considered.

The time-reduced electric and magnetic field wave equations, valid for all linear inhomogeneous media, in terms of the total current density [Stratton, 1941], are, respectively,

$$\nabla \times \nabla \times \bar{E}(\bar{x}) - k_0^2 \bar{E}(\bar{x}) = i\omega\mu_0 \bar{J}(\bar{x}) \quad (28)$$

$$\nabla \times \nabla \times \bar{H}(\bar{x}) - k_0^2 \bar{H}(\bar{x}) = \nabla \times \bar{J}(\bar{x}) \quad (29)$$

which, with the aid of Maxwell's first and second equations, and the equation of continuity for the total charge and current density, can be written as

$$\nabla^2 \bar{E} + k_0^2 \bar{E} = -i\omega\mu_0 \left(\bar{J} + \frac{1}{k_0^2} \nabla \nabla \cdot \bar{J} \right) \quad (30)$$

$$\nabla^2 \bar{H} + k_0^2 \bar{H} = -\nabla \times \bar{J} \quad (31)$$

For nonmagnetic media and perfectly conducting media, the appropriate constitutive equations for the total volume and surface current density $\bar{J}(\bar{x})$ and $\bar{K}(\bar{x})$, respectively, are

$$\bar{J} = (\sigma_f - i\omega\epsilon_e) \bar{E} \quad (32)$$

$$\bar{K} = \hat{n} \times \bar{H} \quad (33)$$

where σ_f and ϵ_e are the free charge conductivity and electric susceptibility, respectively, of a (nonmagnetic) medium and \hat{n} is the outward surface unit vector of a perfectly conducting medium.

Equation (33) is usually regarded as a boundary condition for perfect conductors, but in the context of this paper, it must be taken as a constitutive equation in the truest sense, particularly if regarded as a geometrically constraining condition on the flow of all charges.

For the electric and magnetic field equations, consistent with Eqs. (13) and (26) the appropriate k -space field equations and Green's tensor and vector are, respectively,

$$\bar{E}(\bar{k}) = \Gamma(\bar{k}) \cdot \bar{J}(\bar{k}) + \bar{E}^i(\bar{k}) \quad \Gamma(\bar{k}) \equiv i\omega\mu_0 \Phi(k) \left(I - \frac{1}{k_0^2} \bar{k}\bar{k} \right) \quad (34)$$

$$\bar{H}(\bar{k}) = \bar{\Gamma}(\bar{k}) \times \bar{J}(\bar{k}) + \bar{H}^i(\bar{k}) \quad \bar{\Gamma}(\bar{k}) \equiv i\bar{k} \Phi(\bar{k}) \quad (35)$$

and, by the relationship between surface and volume current densities, consistent with the FFT notation of (19), the volume current density for a perfectly conducting medium, can be written by Eq. (33) as

$$\begin{aligned} \bar{J} &= \frac{ds}{dv} \bar{K} \\ &= \frac{\Delta s}{\Delta v} \hat{n} \times \bar{H} \\ &= \frac{\Delta \bar{s}}{\Delta^3 x} \times \bar{H} \end{aligned} \quad (36)$$

where $\Delta \bar{s}$ is the (finite) differential vector surface area in the FFT cell of volume $\Delta v = \Delta^3 x$.

Equations (32), and (34) and (33), and (35) can now be solved numerically by Eq. (18). Defining the range and phase normalized scattered far-field $\bar{S}(\bar{k}_s)$ as

$$\bar{S}(\bar{k}_s) \equiv \sqrt{4\pi} \lim_{|\bar{x}| \rightarrow \infty} |\bar{x}| \bar{F}^s(\bar{x}) e^{-i\bar{k}_s \cdot \bar{x}} \quad (37)$$

where $\bar{F}^s(\bar{x})$ is the scattered field satisfying the relationship $\bar{F} = \bar{F}^i + \bar{F}^s$, which is consistent with the conventional definition [Berkowitz, 1965a] of the radar power cross section σ and the relationship

$$\sigma = \bar{S} \cdot \bar{S}^* \quad (38)$$

readily reveals that the range and phase normalized electric and magnetic scattered far fields \bar{S}_e and \bar{S}_m are given directly by the k -space current density distributions

$$\bar{S}_e(\bar{k}_s) = \frac{ik_0 z_0}{\sqrt{4\pi}} \left[\bar{J}(\bar{k}) - \widehat{\bar{k}} \widehat{\bar{k}} \cdot \bar{J}(\bar{k}) \right] \Big|_{\bar{k} = \bar{k}_s} \quad (39)$$

$$\bar{S}_m(\bar{k}_s) = \frac{i}{\sqrt{4\pi}} \bar{k} \times \bar{J}(\bar{k}) \Big|_{\bar{k} = \bar{k}_s} \quad (40)$$

where z_0 is the impedance of free space. As dictated by the transversality of the scattered far fields in free space, \bar{S}_e , \bar{S}_m , and \bar{k}_s are indeed all orthogonal to each other.

It can thus be shown that the conventionally defined [Berkowitz, 1965b] electric polarization scattering matrix $\rho_{\eta\xi}$ is given directly in k space by

$$\rho_{\eta\xi} = \frac{ik_0 z_0}{\sqrt{4\pi}} \hat{\mu}_s \cdot \bar{J}(\bar{k}) \Big|_{\bar{k} = \bar{k}_s} \quad (41)$$

where, in the conventional notation for spherical coordinates, $\hat{\eta}_s$ are the eigenpolarizations (spherical coordinate unit base vectors) $\hat{\phi}_s$ and $\hat{\theta}_s$ associated with the scattered far-field propagation vector \bar{k}_s and $\bar{J}(\bar{k})$ is the k -space current density induced by an electric incident plane wave field of the form and polarization

$$\bar{E}^i(\bar{x}) = \hat{\xi} e^{ik_i \cdot \bar{x}} \quad (42)$$

where $\hat{\xi}$ are the eigenpolarizations (spherical coordinate unit base vectors) $\hat{\phi}_i$ and $\hat{\theta}_i$ associated with the incident propagation vector \bar{k}_i .

The solution to the k -space formulation of the three-dimensional scattering problem—for the electric field Eqs. (32) and (34) for nonmagnetic media, and the magnetic field Eqs. (35) and (36) for perfect conductors—has been numerically computer executed for a limited number of cases by the iterative method of solution of Eq. (18), with final results within about one decibel of exact known analytic closed-form solutions after about 30 iterations. Figures 1 through 5 compare this technique with the exact solution of Mie [1908] for the perfectly conducting sphere (of radius a). The failure of this k -space technique in the near vicinity of $k_0 a = 2.75$ (fig. 1) is due to the fact that $k_0 a = 2.75$ is the occurrence of the first eigenfrequency (internal resonance of perfectly spherical shell) [Oshiro et al., 1970]. This difficulty can be readily and simply alleviated, but since the objective of this project was to prove the feasibility and merits of the k -space method, and not the generation of an operational user library of computer programs, this difficulty was taken as beyond the scope of this project.

QUANTUM SCATTERING

Three dimensional nonrelativistic Schrödinger scattering at constant energy by strong finite short-ranged nuclear potentials $V(x)$ is considered. The time-independent Schrödinger equation in center-of-mass coordinates can be written in x space as

$$(\nabla^2 + k_0^2) \psi(\bar{x}) = -\rho(\bar{x}) \quad (43)$$

where

$$k_0^2 \equiv \frac{2mE}{\hbar^2} \quad (44)$$

$$\rho(\bar{x}) \equiv -\frac{2m}{\hbar^2} V(\bar{x}) \psi(\bar{x}) \quad (45)$$

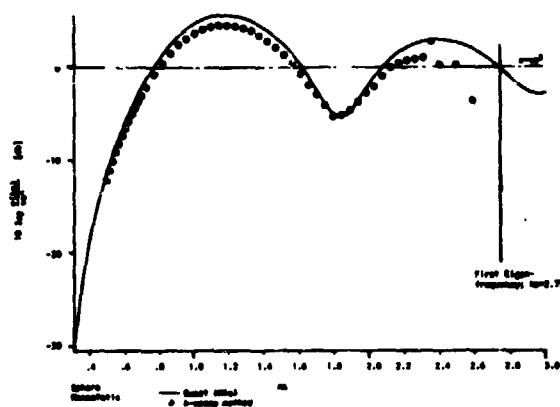


Figure 1. - Power cross section versus Frequency.

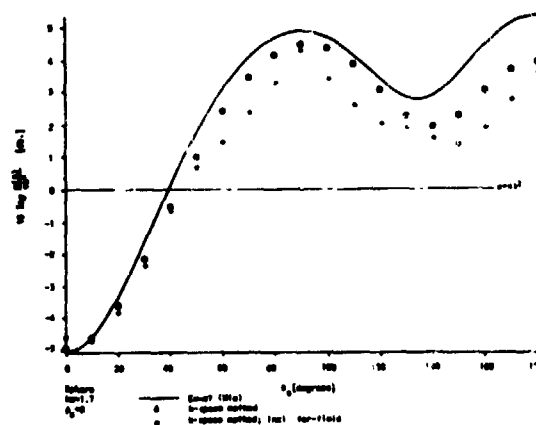


Figure 2. - Power cross section versus aspect angle;
 $k_0 a = 1.7; \phi_0 = 0$

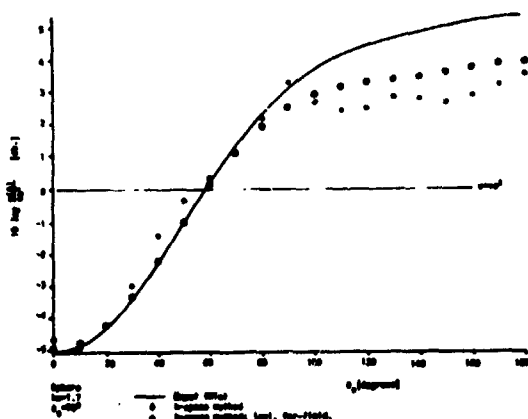


Figure 3. - Power cross section versus aspect angle;
 $k_0 a = 1.7; \phi_0 = 90^\circ$

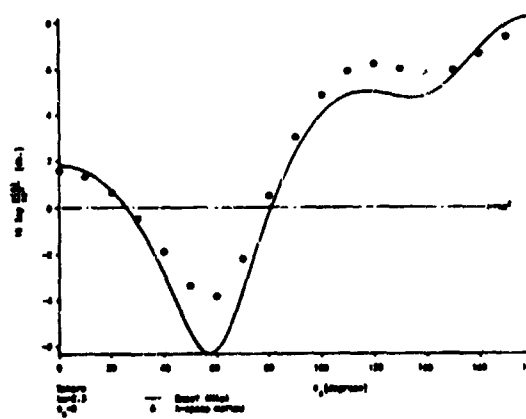


Figure 4. - Power cross section versus aspect angle;
 $k_0 a = 2.3; \phi_0 = 0$

The initial-value integral representation [Rodberg and Thaler, 1967] of which is

$$\psi(\bar{x}) + \frac{2m}{\hbar^2} \int_D \phi(\bar{x}|\bar{x}') v(\bar{x}') \psi(\bar{x}') d^3x' + \psi^I(\bar{x}) \quad (46)$$

Consistent with Eqs. (13) and (26), the appropriate k -space Green's function is thus given by

$$\Phi(\bar{k}) = \frac{i}{k^2 - k_0^2} \leftrightarrow \phi(\bar{x}) = \frac{e^{ik_0|\bar{x}|}}{4\pi|\bar{x}|} \quad (47)$$

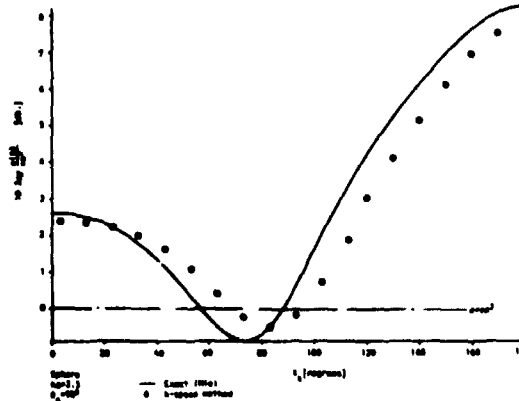


Figure 5. - Power cross section versus aspect angle;
 $k_0 a = 2.3$; $\phi_0 = 90^\circ$

and, consistent with Eqs. (13) and (14), the k -space representation of the initial-value Schrödinger field equation is

$$\Psi(\bar{k}) = \Phi(\bar{k}) \rho(\bar{k}) + \Psi^I(\bar{k}) \quad (48)$$

subject to the x -space constitutive equation

$$\rho(\bar{x}) = -\frac{2m}{\hbar^2} v(\bar{x}) \psi(\bar{x}) \quad (49)$$

Consistent with the inequality (17), the unique existence of this k -space representation is restricted to potentials satisfying the relationship

$$\int |V(x)| d^3x < \infty \quad (50)$$

and the iterative numerical solution (18) is applicable if the potential $V(\bar{x})$ vanishes at a finite distance. The recursion relationship for this iterative solution clearly is

$$\begin{aligned} \rho_n(\bar{k}) &= F(\bar{k}|\bar{x}) \rho_n(\bar{x}) \\ \Psi_n(\bar{k}) &= \Phi(\bar{k}) \rho_n(\bar{k}) + \Psi^I(\bar{k}) \\ \psi_n(\bar{x}) &= F(\bar{x}|\bar{k}) \Psi_n(\bar{k}) \\ \rho_{n+1}(\bar{x}) &= -\frac{2m}{\hbar^2} \alpha v(\bar{x}) \psi_n(\bar{x}) + (1-\alpha) \rho_n(\bar{x}) \end{aligned} \quad (51)$$

Consistent with Eqs. (37) through (40), the range- and phase-normalized scattered far-field wave functions $S(\bar{k}_s)$ are yielded directly in k -space as

$$S(\bar{k}) = \frac{1}{\sqrt{4\pi}} \rho(\bar{k}) \Big|_{\bar{k}=\bar{k}_s} \quad |\bar{k}_s| = \bar{k}_0 \quad (52)$$

and the differential and total scattering cross sections σ_d and σ_t are

$$\sigma_d(\bar{k}_s) = \bar{S}(\bar{k}_s) \cdot \bar{S}^*(\bar{k}_s) \quad (53)$$

$$\sigma_t = k_0^{-2} \oint_{|\bar{k}_s|=k_0} \sigma_d(\bar{k}) \cdot da_k \quad (54)$$

where da_k is the area of the spherical surface in k space of radius k_0 .

If the potential is a nonlocal potential representable by a difference kernel — for example,

$$\rho(\bar{x}) = -\frac{2m}{\hbar^2} \int v(\bar{x}-\bar{x}') \psi(\bar{x}') d^3\bar{x}' \quad (55)$$

the constitutive Eq. (49) becomes a local equation in k space:

$$P(\bar{k}) = -\frac{2m}{\hbar^2} V(\bar{k}) \Psi(\bar{k}) \quad (56)$$

and the scattering problem represented now by Eqs. (48) and (56) can be solved in closed form in k space:

$$\Psi(\bar{k}) = -\frac{2m}{\hbar^2} V(\bar{k}) \Psi(\bar{k}) \Phi(\bar{k}) + \Psi^i(\bar{k}) \quad (57)$$

$$\Psi(\bar{k}) = \frac{\Psi^i(\bar{k})}{1 + \frac{2m}{\hbar^2} V(\bar{k}) \Phi(\bar{k})} \quad (58)$$

ACKNOWLEDGMENT

The author is indebted to Mr. Charles H. Krueger, Jr., of the Air Force Avionics Laboratory, Wright-Patterson Air Force Base, Ohio, for developing the details, programming and verifying, the two-dimensional k -space formulation of the electromagnetic scattering problem. This author is also indebted to Messrs. Leslie E. Whitford and William Hopkins of the Computing and Information Systems Division, Computer Science Center, Wright-Patterson Air Force Base, Ohio, for directing the programming effort and performing the actual programming, respectively, of the one- and three-dimensional k -space formulations of the electromagnetic scattering problem, as well as for several important suggestions.

This research was supported in part by U.S.A.F. contract F 33615-70-C-1345.

DISCUSSION

Unidentified speaker: What do you mean by the "fast" matrix form? You mean the fact that you can do it quickly?

Bojarski: Well, that is a good way of describing it. It is basically a Fourier transform. If you write Eq. (10) and if you compute it numerically and it were to go straightforward it would take you n^2 operations to obtain n values. But if you cleverly use the periodicity, which was done at the turn of the century but has only been programmed recently, you will find out that you don't have to compute n^2 operations but only $n \log n$. By the way, we did do it in three dimensions, so you can do three-dimensional problems and larger.

Kay: I think that this is a very clever use of the fast Fourier transform, and I think maybe something could be done, but I can see a fundamental difficulty. Your second equation is an integral equation which is a convolution, and therefore could theoretically be inverted.

Bojarski: Let me rewrite it so we will know what we are talking about. The first one is Eq. (13) and the second one is Eq. (14).

Kay: Okay. Now, in the first one, $f(k)$ is the Fourier transform of the product of two functions of x . Write it out in terms of the unknown function $\sigma(x)$ as a convolution, an integral equation of the convolution form that can be inverted by the Fourier transform.

Bojarski: Yes, but that is precisely the thing to avoid, to put it in the convolution form.

Kay: I understand that, but I am just commenting that the success of the fast Fourier transform seems to depend on this fact, and the problem with the Gelfand-Levitan methods in the inverse problem is that the kernel of the integral equation has the right form but it isn't one that can be inverted by the Fourier transform, so it is not quite clear that this technique would work in that case, for example.

Bojarski: Perhaps we shouldn't go to that but go back to the original differential equations. I needn't have written it for this operator. In fact, for the electromagnetic case it is of the form $\text{del cross del cross}$, which you put in other forms. Basically this technique will work so long as you split apart a physical law that has a differential operator on a field, giving you a source, and a constitutive equation that says the source is related to the field. As long as this is a linear differential equation the technique will work.

Kay: Yes, but it does depend on using a Fourier transform, so this means that the equivalent integral equation has to be one that can be inverted by the Fourier transform. I am not posing this as a fundamental objection. It is just that I can see a difficulty arising because of that question.

REFERENCES

1. Harrington, R. F.: *Field Computation by Moment Methods*, Macmillan, 1968
2. Papoulis, A.: *The Fourier Integral and its Applications*, McGraw-Hill, 1962.
3. Cooley, J. W.; Lewis, P. A. W.; and Welch, P.D.: *The Fast Fourier Transform Algorithm and its Applications*, IBM Research Paper 1743, Sect. 10, July 1967.
4. Varga, R. S.: *Matrix Iterative Analysis*, Sect. 3 et seq., Prentice-Hall, 1962.
5. Cooley, J. W., et al.: *Op. Cit.*
6. Van Bladel, J.: Some Remarks on the Green's Dyadic for Infinite Spaces. *IRE Transactions on Antennas and Propagation*, vol 9, Nov. 1961, pp. 563-566, eq. 5.
7. Singleton, R. C.: An Algorithm for Computing the Mixed Radix Fast Fourier Transform. *IEEE Trans. on Audio and Electroacoustics*, vol. AU-17, June 1969, pp 93-103 (p. 95).
8. Singleton, R. C.: Personal Communication, April 1970.
9. Stratton, J. A.: *Electromagnetic Theory*, Sect. 1.5, p. 12, McGraw Hill, 1941.
10. Berkowitz, R. S.: *Modern Radar*, Sect. 5.3, p. 550, Wiley, 1965a.
11. *Ibid*, Sect. 5.5, pp. 560-565, 1965b.
12. Mie, *Ann. Physik*, vol. 25, p. 377, 1908.
13. Oshiro, F. K.; Mitzner, K. M.; Locus, S. S.; et al.: *Calculation of Radar Cross Section*. Northrop Corp. Technical Report AFAL-TR-70-21, Part II, Sect. V, p. 109 et seq., April 1970.
14. Rodberg, L. S.; and Thaler, R. M.: *Introduction to the Quantum Theory of Scattering*, Sect. 5, p. 127 et seq., Academic Press, 1967.

INVERSION OF RADIATION DATA IN BIOPHYSICS

Victor Twersky

University of Illinois

1 N73-11624

This summarizes the essentials of my informal supplementary talk during the panel discussion (final workshop session) and incorporates some remarks relevant to comments from the floor.

There are essentially two sets of topics in biophysics that are appropriate. The first relates to how we acquire information about our distant environment through seeing, hearing, and so forth. The second relates to how we use electromagnetic, acoustic, or other radiation for diagnostic purposes, either at a bulk or a molecular level.

The first set includes inversion aspects not covered by this conference, in that the mapping characteristics of the receiver (the sensor-perceptory system) are as important as the scatterers. An elegant example, *Luneburg's* [1948, 1950] theory of binocular vision [*Blank*, 1953, 1961] was motivated by work on stereoptical and optical localization equipment (such as long-arm binoculars to widen the base for triangulation). *Luneburg* was able to account for various results in the literature of experimental psychology, particularly for the noncoincidence of parallel alley and equidistant alley test data, by a model based on non-Euclidean geometry: under restricted conditions, the subject exposed to optical stimuli in euclidean three-space maps them into a perceptual three-space of constant negative curvature.

As another example, consider the development of acoustic guidance devices to aid the blind in foot travel [*Twersky*, 1947, 1951a,b], such as hand-held high-frequency sound sources to supplement or improve on ambient noise (or heel clicks, etc.) for obstacle avoidance by hearing. The average increase in scattered sound intensity at the ear as an obstacle is approached provides two kinds of auditory cues: because of the nonlinear behavior of the hearing mechanism, an increase in intensity of the high-frequency stimulus is perceived as an increase in pitch as well as loudness.

The second set of relevant topics, radiation diagnostics, involves the same fundamental physical processes that motivated the present conference. Thus, the mathematical developments that have been presented are as significant to scattering phenomena in biophysics as in geophysics, quantum physics, etc. The inversion of x-ray, optic, or ultrasonic scattering and propagation data to determine the density profile of tumors and of tissue layers, the shape of foreign objects in the body, the velocity distributions in vital fluid systems, and the like [*Medical Physics*, 1944, 1950, 1960] provide full analogs of the data inversion problems already considered.

In addition, there are various many-obstacle problems, periodic arrays, or statistical distributions, that are of interest. Although a many-obstacle configuration can be thought of as one scatterer in disjoint pieces (so that the analytical procedures already discussed are applicable), it is usually more efficient to exploit the known results for the components as isolated scatterers. We represent the many-obstacle solution (for nonpenetrating scattering potentials) as a functional of isolated-scatterer functions, and concentrate on the novel aspects arising from the configuration [*Twersky*, 1952; 1962a,b; 1967a; *Burke et al.*, 1965] of scatterer locations, or their statistical distribution [*Foldy*, 1945, 1951, 1952; *Keller*, 1964; *Twersky*, 1959, 1962c,d, 1964, 1969].

A well-known example of a biologically significant configuration discovered *in part* by interpreting x-ray scattering data [*Langridge et al.*, 1960] is the double helix of DNA. (The "part" is stressed because much of the information used by *Watson and Crick* come from direct chemical analysis of the ratios of the four bases that distinguish the nucleotides, from electrometric titration, from viscosity studies, and so forth; but one does not play a gentleman's game with nature.) The x-ray data are obtained from measurements on a relatively thick fiber consisting of billions of aligned DNA molecules. In the plane perpendicular to the fiber's axis, the distribution of the helices should be essentially that of a two-dimensional liquid; thus, the fiber data may show additional structure, analogous to that shown by x-ray data for dense gases and liquids [*Guinier*, 1963], from which the two-dimensional pair-distribution function could be obtained by inversion.

A single-scattering approximation for the nucleotides suffices for x-ray studies on DNA. However, for ultraviolet studies, the interactions of the fields of the nucleotides must be taken into account. As discussed by Weissbluth [1971], the ultraviolet attenuation is smaller for a solution of DNA helices than for a solution of random coils (as obtained, for example, by heating), because the applied field couples the oscillator strengths (electric dipoles) of neighboring nucleotides in a helix.

As a final example we consider the inversion of optical data to determine the oxygen saturation R of blood: the ratio R of oxygenated to total hemoglobin is one of the principal parameters of interest in clinical medicine. If a sample of blood is hemolyzed — that is, the hemoglobin is released from the red blood cells — then R for the hemoglobin solution can be determined by optical absorption measurements at two different wavelengths. (This is essentially a refined version of observing the color of blood, bright red for high R as in arterial blood, and dark red for low R as in venous blood.) To avoid sampling, optical instruments (transmission oximeters for the shell of the ear, reflection oximeters for the forehead, fiber-optical catheters for insertion in blood vessels) and empirical calibration procedures have been developed for continuously monitoring R during anesthesia, spaceflights, etc. To design more versatile instruments, the inversion procedure [Loewinger *et al.*, 1969; Anderson and Sekelj, 1967a,b] must take account of the scattering effects [Twersky, 1970] of the large red blood cells in which the hemoglobin molecules are packaged. For transmission oximetry on suspensions of whole red blood cells, the main beam is attenuated by scattering to other directions as well as by molecular absorption; in addition, there are the enrichment effects of scattering into the beam. For a reflection system, incoherent scattering interface effects are also significant [Twersky, 1970].

In connection with comments on measuring phase as well as amplitude, it should be noted that various parameters sought by inverting the total average intensity and coherent phase shift could be obtained more directly from averages of several products of the phase quadrature components of the field [Twersky, 1963, 1967b; Beard *et al.*, 1967; Hawley *et al.*, 1967]. The components may be measured with a coherent source and a double interferometer analogous to the double-bridge heterodyne system developed for millimeter wave measurements on scattering by large-scale random distributions [Twersky, 1963, 1967b; Beard *et al.*, 1967; Hawley *et al.*, 1967]. Such systems enable us to separate the coherent and incoherent intensities even in the forward direction, to measure new combinations of the fundamental parameters, and to seek data in which the many-particle distribution functions enter in first order instead of as small corrections.

REFERENCES

1. Anderson, N. M.; and Sekelj, P. S.: Light-Absorbing and Scattering Properties of Nonhaemolyzed Blood. *Phys. Med. Biol.*, vol. 12, 1967a, pp. 173-184.
2. Anderson, N. M.; and Selelj, P. S.: *Reflection and Transmission of Light by Thin Films of Nonhaemolyzed Blood*, 1967b, pp. 185-192.
3. Beard, C. I.; Kays, T. H.; and Twersky, V.: Scattering by Random Distributions of Spheres Versus Concentration. *IEEE Trans.* AP-15, 1967, pp. 99-118.
4. Blank, A. A.: Curvature of Binocular Visual Space. *J. Opt. Soc. Amer.*, vol. 51, 1961, pp. 335-339.
5. Blank, A. A.: The Luneburg Theory of Binocular Visual Space, *J. Opt. Soc. Amer.*, vol. 43, 1953, pp. 717-727.
6. Burke, J. E.; Censor, D.; and Twersky, V.: Exact Inverse-Separation Series for Multiple Scattering in Two Dimensions. *J. Acoust. Soc. Amer.*, vol. 37, 1965, pp. 5-13.
7. Foldy, L. L.: The Multiple Scattering of Waves. *Phys. Rev.*, vol. 67, 1945, pp. 107-112.
8. Glasser, O., ed.: *Medical Physics*, vol. I, 1944; vol. II, 1950; vol. III, 1960, Year Book Pub., Chicago (contains review articles on diffraction microscopes, Roentgenrays, etc., with applications to diagnostics).
9. Guinier, A.: *X-Ray Diffraction*, W. H. Freeman and Co., San Francisco, 1963.
10. Hawley, S. W.; Kays, T. H.; and Twersky, V.: Comparison of Distribution Functions from Scattering Data on Different Sets of Spheres. *IEEE Trans.*, AP-15, 1967, pp. 118-135.
11. Keller, J. B.: Stochastic Equations and Wave Propagation in Random Media. *Proc. Symp. Appl. Math.*, vol. 16, Amer. Math. Soc., Providence, R. I., pp. 145-170.
12. Langridge, R., et al.: The Molecular Configuration of DNA. *J. Mol. Biol.*, vol. 2, 1960, pp. 19-64.
13. Lax, M.: Multiple Scattering of Waves, I. *Rev. Mod. Phys.*, vol. 23, 1951a, pp. 287-310.
14. Lax, M.: II, The Effective Field in Dense Systems. *Phys. Rev.*, vol. 88, 1952b, pp. 621-629.
15. Loewinger, E.; Gordon, A.; Weinreb, A.; and Gross, J.: Analysis of a Micromethod for Transmission Oximetry of Whole Blood. *J. Appl. Physiol.*, vol. 19, 1964, pp. 1179-1184.
16. Luneburg, R. K.: *Mathematical Analysis of Binocular Vision*, Princeton Univ. Press, 1947.
17. Luneburg, R. K.: Metric Methods in Binocular Visual Perception. *Studies and Essays: Courant Anniversary Volume*, Interscience Pub., N. Y., 1948, pp. 215-240.
18. Luneburg, R. K.: The Metric of Binocular Visual Space, *J. Opt. Soc. Amer.*, vol. 40, 1950, pp. 627-642.
19. Twersky, V.: An Obstacle Detecting Device for the Blind. *Biol. Rev.*, vol. 9, 1947, pp. 16-21.
20. Twersky, V.: Flashsounds and Aural Constructs for the Blind. *Physics Today*, vol. 4, 1951a, pp. 10-16.

21. Twersky, V.: On the Physical Basis of the Perception of Obstacles by the Blind. *Amer. J. Psychol.*, vol. 64, 1951b, pp. 409-416.
22. Twersky, V.: Multiple Scattering of Radiation by an Arbitrary Planar Configuration of Parallel Cylinders and by Two Parallel Cylinders. *J. Appl. Phys.*, vol. 23, 1952a, pp. 407-414.
23. Twersky, V.: Scattering by Quasi-Periodic and Quasi-Random Distributions. *IRE Trans.*, AP-7, 1959, S307-S319.
24. Twersky, V.: Scattering of Waves by Two Objects. *Electromagnetic Waves*, edited by R. E. Langer, Univ. Wisconsin Press, 1962a.
25. Twersky, V.: Multiple Scattering by Arbitrary Configurations in Three Dimensions. *J. Math. Phys.*, vol. 3, 1962b, pp. 83-91.
26. Twersky, V.: On Scattering of Waves by Random Distributions. *J. Math. Phys.*, vol. 3, 1962c, pp. 700-715, 724-734.
27. Twersky, V.: Multiple Scattering of Waves and Optical Phenomena. *J. Opt. Soc. Amer.*, vol. 52, 1962d, pp. 145-171.
28. Twersky, V.: Signals, Scatterers and Statistics. *IEEE Trans.*, AP-11, 1963, pp. 668-680.
29. Twersky, V.: On Propagation in Random Media of Discrete Scatterers. *Proc. Symp. Appl. Math.*, vol. 16, Amer. Math. Soc., Providence, R. I., 1964, pp. 84-116.
30. Twersky, V.: Multiple Scattering of Electromagnetic Waves by Arbitrary Configurations. *J. Math. Phys.*, vol. 8, 1967a, pp. 589-610.
31. Twersky, V.: Theory and Microwave Measurements of Higher Statistical Moments of Randomly Scattered Fields. *Electromagnetic Scattering*, edited by R. L. Rowell and R. S. Stein, Gordon and Breach, N. Y., 1967b, pp. 579-695.
32. Twersky, V.: Scattering by Discrete Random Media. *Turbulence of Fluids and Plasmas*, edited by J. Fox, Polytechnic Press, Brooklyn, N. Y., 1969, pp. 143-161.
33. Twersky, V.: Absorption and Multiple Scattering by Biological Suspensions. *J. Opt. Soc. Amer.*, vol. 60, 1970, pp. 1084-1093.
34. Weissbluth, M.: Hypochromism. *Quart. Rev. Biophys.*, vol. 4, 1971, pp. 1-34.

TRANSMISSION LINE INVERSION AND SYNTHESIS FROM THE POINT OF VIEW OF TRANSIENT RESPONSE

M. M. Sondhi

Bell Telephone Laboratories, Inc., Murray Hill, New Jersey

N73-11625

The work described here is a joint effort of *Sondhi and Gopinath* [1971a,b; *Gopinath and Sondhi*, 1971].

Although the problem is discussed as one of transmission line inversion (or synthesis), the motivation for studying it comes from research in speech communication. Since this field of research is rather remote from the other disciplines represented here, we first briefly consider the problem in its original form and then reduce it to one of transmission lines, indicating the most recent results obtained.

Figure 1 is a cross section of the human vocal tract. To produce speech sounds we acoustically excite this tract by means of quasi-periodic pulses from the vocal cords during vowel sounds and turbulent air flow at various points along the tract during fricative sounds such as s, sh, and f. If we assume plane wave propagation (which is a good approximation for propagation of waveforms limited to about 3500 Hz), then the cross-sectional area $C(x)$, the pressure $p(x, t)$, and the volume velocity $u(x, t)$, which is defined as $C(x)$ times the particle velocity, are related through the differential equations

$$\begin{aligned} C \frac{\partial p}{\partial t} &= - \frac{\partial u}{\partial x} \\ \frac{1}{C} \frac{\partial u}{\partial t} &= - \frac{\partial p}{\partial x} \end{aligned} \quad (1)$$

where the units are chosen so that the density of air and the velocity of sound are both unity. We identify $x = 0$ with the lips and make the further normalization $C(0) = 1$ — that is, the lip opening has unit area. These three normalizations completely specify a mass-length-time system of units.

Now if the shape of the tract is known — that is, in our idealization to plane waves, $C(x)$ is known — and the excitation is known, we can compute the pressure at the lips — the speech wave. This computation involves solution of Eqs. (1) for given boundary conditions.

The inverse problem of interest here is to infer $C(x)$ from acoustic measurements at the lips. That is the sort of information one requires for synthesis of speech, making models of the speech production mechanism, and so forth.

The problem becomes a transmission line problem if we replace the pressure $p(x, t)$ by the voltage $v(x, t)$ and the volume velocity $u(x, t)$ by the current $i(x, t)$. With these identifications Eqs. (1) become

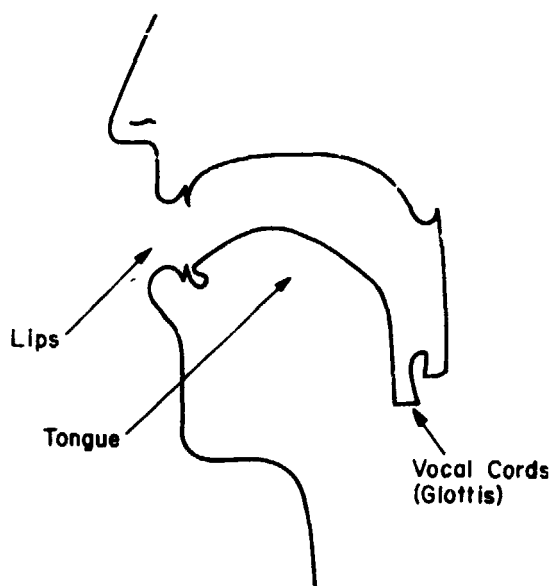


Figure 1.

$$C \frac{\partial v}{\partial t} = - \frac{\partial i}{\partial x} \quad (2)$$

$$\frac{1}{C} \frac{\partial i}{\partial t} = - \frac{\partial v}{\partial x}$$

which are the equations discussed by Kay (Chap. 6) except that he used the steady-state, frequency domain version with $\partial/\partial t$ replaced by $i\omega$. In fact, Eqs. (2) are the equations for a special type of line for which $C(x)$ is the capacitance per unit length and also the reciprocal of the inductance per unit length. This reciprocal relation between capacitance and inductance eliminates the inherent ambiguity of the general transmission line inversion problem. Recall the fact brought out by Kay that input measurements alone *cannot* yield both the inductance and the capacitance of an arbitrary line. Only one function derived from these quantities can be reconstructed. This function may be chosen as, for example, the characteristic impedance as a function of electrical distance or some other equivalent form. In our case, because of the reciprocal relation the function to be reconstructed may be chosen to be $C(x)$ itself.

The end $x = 0$ of this transmission line is accessible to us for measurement. As the Eqs. (2) are linear it is clear that if we can measure the impulse response at the input then we have measured all we possible can. All other measurements can be predicted from the impulse response by superposition. The impulse response is defined here to be the voltage $h(t)$ developed at $x = 0$, t units of time after a unit impulse of current is applied to the quiescent line at $x = 0$. It can be shown that $h(t) = \delta(t) + \hat{h}(t)$ where $\hat{h}(t)$ is continuous if $C(x)$ is continuously differentiable. (We assume this degree of smoothness for $C(x)$, although this is not necessary.)

With this definition of the impulse response if the transmission line is quiescent at $t = t_0$ — that is, $v(x, t_0) = i(x, t_0) = 0$ for all $x > 0$ — then the input voltage and current are related by the convolution

$$v(0, t) = i(0, t) + \int_{t_0}^t \hat{h}(t - \tau) i(0, \tau) d\tau$$

Our problem may now be stated precisely as follows: We are given a transmission line, described by Eqs. (2), with $C(x)$ continuously differentiable and $0 < C(x) < \infty$. At $x = 0$ we measure $h(t)$ for the interval $0 \leq t \leq T$. Can this measurement tell us anything about $C(x)$?

Since the velocity of propagation is unity, it is clear that $h(t)$ for $t < T$ is unaffected by the properties of the line for $x > T/2$. Those remote regions of the line just cannot be probed in so short a time from the input. Thus $h(t)$ for $t < T$ tells us *nothing* about $C(x)$ for $x > T/2$. The remarkable thing, however, is that it gives $C(x)$ uniquely for $x < T/2$.

The basic result is as follows: With $h(t)$ given for $t < T$ choose any $\alpha < T/2$; solve the integral equation

$$f(t) + \frac{1}{2} \int_{-\alpha}^{\alpha} \hat{h}(|t - \tau|) f(\tau) d\tau = 1 \quad |t| < \alpha \quad (3)$$

for the unknown function $f(t)$. Then

$$\int_{-\alpha}^0 f(t) dt = \int_0^{\alpha} f(t) dt = \int_0^{\alpha} C(x) dx \quad (4)$$

which solves the problem.

The argument leading to this result is simple and is based on the notions of causality and charge conservation. Suppose the line is quiescent at $t = -\alpha$, at which time a current $i(0, t)$ starts to flow in at $x = 0$. Causality and unit propagation velocity then assure us that at $t = 0$ the line will be undisturbed for $x > \alpha$. Thus, the charge delivered by the input current must appear somewhere on the portion $x \leq \alpha$ of the line. From charge conservation, therefore,

$$\int_{-\alpha}^0 i(0, t) dt = \int_0^{\alpha} v(x, 0) C(x) dx \quad (5)$$

What one can show is that if $i(0, t) = f(t)$, where $f(t)$ is the solution of Eq. (3) then $v(x, 0) = 1$ for all $x < \alpha$. Thus, we obtain Eq. (4).

To obtain $C(x)$ one has to differentiate the integrated capacitance obtained from Eq. (4). For numerical computation this is a very undesirable operation. It turns out to be unnecessary, for a simple manipulation of Eqs. (3) and (4) shows that

$$f^2(\alpha) = C(\alpha) \quad (6)$$

Note, however, that differentiation of Eq. (4) with respect to α does not give $f(\alpha) = C(\alpha)$ because $f(t)$ in that equation depends on α implicitly.

The impulse response is a positive definite operator; mathematically, this means that the inequality

$$\int_{-\alpha}^{\alpha} g^2(t) dt + \frac{1}{2} \int_{-\alpha}^{\alpha} \int_{-\alpha}^{\alpha} \hat{h}(|t-\tau|) g(t) g(\tau) d\tau > 0$$

holds for any square integrable function $g(t)$ on the interval $|t| \leq \alpha \leq T/2$. Physically, this inequality states that an input current $g(t)$ delivers a positive amount of energy to the line because of the passivity of the line and because $C(x) \neq 0$ or ∞ . This positive definite property ensures the existence and uniqueness of the solution of Eq. (3).

In conclusion it should be noted that if $h(t)$ is specified a priori rather than measured, we have a synthesis problem. The theory, of course, applies equally well to this case provided the specified function is a valid impulse response. Clearly $h(t)$ must satisfy the two necessary conditions discussed earlier — that is, it must be positive definite and of the form $\delta(t) + \hat{h}(t)$. We can show that these conditions are also sufficient to construct $C(x)$. At least if $\hat{h}(t)$ is continuous and continuously differentiable $C(x)$ the theory is complete in the sense in which Sabatier used the term (Chap. 5). That is to say the relation $h(t) \leftrightarrow C(x)$ is unique in each direction.

DISCUSSION

Moses: I want to say I am very much impressed with this paper, and I will tell you some reasons why. First of all, I want to correct an impression. The method of Gelfand-Levitan, the way Irv Kay does it does hold for time, so it can be put in the time-dependent form. You can send in a time-dependent impulse and get something out. There is no difficulty whatever. In fact, he presented it in the time-dependent form, actually, when he did it through causality, with a δ function and so on. He immediately transformed, but it is completely unnecessary to transform.

Sondhi: Yes, but the viewpoint has always been in terms of steady-state behavior as far as I know.

Moses: You don't have to do it that way, though, with Gelfand-Levitan.

Sondhi: Perhaps not. Only, they have never pursued it this way.

Moses: Well, in Irv's paper, it will be this way. He has published elsewhere, as a matter of fact. I have to give credit to Irv for recognizing many years ago that the Gelfand-Levitan equation was an aspect of causality, just as this is. Now, you seem to have a different aspect of causality. This is what makes it very interesting.

When Irv and I worked on the three-dimensional Gelfand-Levitan theory we tried to derive the causality and got absolutely stuck, and we couldn't do it. We had to go to the spectral theory. But I think causality is the proper way to do all these problems.

Sondhi: Well, I would like to make one additional point as far as causality is concerned. For the specific problem that we were interested in, if you use the frequency approach, you must know first of all the end condition at the glottis. Now, this end condition is never really known. You could assume that the tube is closed at that end, but this assumption would not always be valid.

Our method doesn't require any information about that. It builds up the cross-sectional area function starting from the end at which you are making measurements, as far down as you can go, depending on how many seconds of the impulse response you have.

Parry: For this particular application, wouldn't it be more natural to start the signal at one end and observe it at the other end?

Sondhi: But that is what we are trying to get around, because we can't put a known signal into the glottis. That's the problem. Ideally, what you would like to do is to get the area function directly from the speech, but you don't know what the excitation is. You can only make some reasonable assumptions about it. However, that might produce a large error in the reconstructed area.

Now the measurement problem is not easy because you have to ask the person to mouth the words without actually speaking. While you are making the measurement you don't want the speech to interfere with it.

Many linguists don't like this, because they say that you are not measuring the shapes you would produce in normal circumstances. How important this effect is no one really knows. However, in measuring the area function no assumptions are involved about the excitation and the boundary conditions.

Parry: As I understood it, you assumed that the cross-sectional area at $x = 0$ was 1? So your results, then, will be given in terms of a ratio of $C(x)$ to $C(0)$? Anyone who has a big mouth, so to speak, would therefore upset your calculations, or, being more serious about it, the point is that the C is not constant at $x = 0$.

Sondhi: Right. We scale down big mouths in the following way. To make this measurement it is not feasible to just put a pulse of sound at the lips and make a pressure measurement right there. What we do is we have what is known as an impedance tube, into which we put some energy, and the person holds his mouth at one end by a flexible coupling. We start reconstructing the area from some point in the tube where we know what the area is. That is the way we get around the problem. But in general you are right. If you were going to make measurements without such a tube, then the scale factor would have to be restored by other means.

Portinari: Did you look into conditions of h such that $C(x)$ is of finite range, i.e., is constant after a finite length?

Sondhi: You mean that $C(x)$ should have some variation and then stay constant?

Portinari: That's right.

Sondhi: I haven't looked at that problem, but if that were the case, you see, it would be easier to look at it as a termination of the transmission line, and solve the problem that way. The uniform tube is equivalent to a resistance termination on a transmission line, so I could look at this variable transmission line with a resistor termination. I am not sure I understood your problem correctly.

Portinari: What I am asking is more of a synthesis question than diagnosis. I mean, you have h . You want to construct $C(x)$, and you would be interested in having something that occupies a finite length. What would be conditions on h to guarantee that?

Sondhi: I have not looked at that problem. We have looked a little bit at transmission lines that include distributed resistive and conductive losses. This is the question I asked Professor Kay, in fact. What happens if you know the resistance and conductance? Can you get C from such measurements?

We don't know the answer in general, but in some special cases we have actually solved that problem, also I have never seen anybody treat the inverse problem including loss in transmission lines before, even in special cases [*Sondhi and Gopinath, 1971 b*].

Mitra: I wonder if Dr. Moses and you would comment on the correspondence between the time domain Gelfand-Levitan equation and this one?

Moses: I haven't studied this equation long enough, really, to make a contrast. With the Gelfand-Levitan, when you do it in the time domain, in which you start at $t = 0$, you assume a time cut-off. Your line doesn't go to infinity. I'll do it in the Schrödinger equation.

You start at $t = 0$. You send the signal in. It bounces back from the reflection coefficient, or the convolution of the reflection coefficient, whatever wave you get back. You can get the reflection coefficient, and you carry out the Gelfand-Levitan. There's a finite time for this packet to come out. The cutoff of the potential permits you to send something else. I assume the same thing happens here.

Mitra: He is making distinctions so he never has to worry about the terminations. He is steady on the line. He is just generating the assumptions.

Shmoys: It is an infinite line, because he considers transient response for a finite time lag.

Sondhi: What Professor Shmoys is saying is right. In T units of time you can only traverse $T/2$ units of the line; whether it is infinite from there on or not is immaterial.

But I don't really know what is meant by the time domain Gelfand-Levitan equation because I have never seen it. I would like to.

Moses: I think causality is the essential feature. I think spectral theory may be a way of solving it, but the causality makes it very interesting.

Unidentified speaker: Harry, you should make that clear, it is the Marchenko version that you have in mind.

Moses: I was thinking, actually, of the Gelfand-Levitan, with 0 to ∞ , I mean $\pm\infty$. You just cut off the potential at 0. Marchenko is very close to that.

REFERENCES

1. Gopinath, B.; and Sondhi, M. M.: Inversion of the Telegraph Equations and the Synthesis of Nonuniform Lines. *IEEE*, vol. 59, No. 3, March 1971, p. 383.
2. Sondhi, M. M.; and Gopinath, B.: Determination of Vocal-Tract Shape from Impulse Response at the Lips. *Acoust. Soc. Amer.*, vol. 49, No. 6 (Part 2) June 1971a, p. 1867.
3. Sondhi, M. M.; and Gopinath, B.: Determination of the Shape of a Lossy Vocal Tract. No. 23 C 10, 7th *Int. Cong. Acoustics, Budapest, Hungary*, August 1971b.

THE INVERSION OF VLF-LF SOUNDER DATA TO OBTAIN D-REGION ELECTRON-DENSITY DISTRIBUTIONS

C. H. Shellman

Naval Electronics Laboratory Center
San Diego, California

N73-11626

ABSTRACT

Work is being carried out at the Naval Electronics Laboratory Center to determine *D*-region electron-density distributions from VLF-LF sounder data. The direct problem is to compute reflection coefficients for an assumed electron-density profile, using a full-wave solution. The inverse problem is treated numerically, using derivatives of the reflection matrix with respect to model parameters. A technique for the rapid computation of the derivatives is described. An algorithm for determining resolution in the profile is discussed.

INTRODUCTION

Our work at the Naval Electronics Laboratory Center is that of finding *D*-region electron-density distributions from very low frequency reflection data. It is an inverse problem, since the direct problem is that of computing reflection coefficients when the electron density profile is known. The direct problem involves the numerical solution of the full-wave problem for a lossy anisotropic ionosphere, since, in the VLF region of the radio spectrum, ray theory no longer applies. Data were first taken in 1961. In an effort to perform the inverse problem of finding electron density profiles from the data, we first attempted trial-and-error methods, rather unsuccessfully.

Function optimization techniques were later used for finding these distributions. *Sheddy* [1963a] has discussed this subject, and an error analysis for relating the uncertainties in the data to uncertainties in the profile subsequently was given by *Shellman* [1969, 1970].

Not many sets of data were analyzed, however, because of the large amount of computation required. The problem can be appreciated a little more by mentioning that it takes about 1 or 2 sec on an IBM 360-65 computer to compute the full-wave solution for one frequency. Data may be taken at as many as ten frequencies, so carrying out a complete set of full-wave solutions for very many trial profiles could run up a computer bill very quickly.

About two months ago we discovered a rapid way of differentiating the numerical integration with respect to electron-density model parameters, which has speeded up our computation considerably. The method is discussed a bit later.

This meeting has certainly demonstrated the importance of interdisciplinary communication, as we have seen parallels with our efforts at NELC, especially in the presentations given by the geophysicists. Doubtless our work with the VLF sounder data could have been much easier had we known of these other efforts sooner.

The region of the ionosphere that we are interested in, namely, the *D*-region, is inaccessible to HF sounders. The densities are generally below 10^6 electrons per cm^3 , and the region was generally denoted by a dotted line at the bottom of the profile in the talks on HF sounders.

The NELC VLF sounder [Paulson and Thelsen, 1962] is located at Sentinel, Arizona, in a region of lava flow, and hence low electrical conductivity. The transmitter (dipole) antenna is oriented along the magnetic north-south direction, and is about 9 km long. The receiver vans are usually placed about 15 km magnetic west of the transmitter. Two receiver loops, oriented perpendicular to one another, are used. One receives the normally reflected skywave and the other receives the conversion component, which is present only because of anisotropy — the influence of the earth's magnetic field. The two components are usually of the same order of magnitude.

Reflection from the ionosphere is described by the reflection matrix

$$R = \begin{bmatrix} |_{\parallel} R_{\parallel} | \exp(i_{\parallel} \phi_{\parallel}) & |_{\perp} R_{\parallel} | \exp(i_{\perp} \phi_{\parallel}) \\ |_{\parallel} R_{\perp} | \exp(i_{\parallel} \phi_{\perp}) & |_{\perp} R_{\perp} | \exp(i_{\perp} \phi_{\perp}) \end{bmatrix}$$

where the first subscript refers to the polarization of the upgoing wave with respect to the plane of incidence and the second subscript refers to the downcoming wave. Because of the orientation of transmitter and receiver, the right half of this matrix applies to the sounder, hence we have four (real) measurables for each sounder frequency.

For convenience of notation, we denote all the observables by R_i and assign to each one an expected standard error σ_{R_i} .

What we are trying to accomplish is the following: We have data from which we want to find an electron density distribution: because the data have uncertainty, and we have a finite amount of data, the density distribution similarly has uncertainty. So we have in the profile a certain resolution and certain error width, the resolution being related mostly to the amount of data, and the error band to both the amount and accuracy of the data.

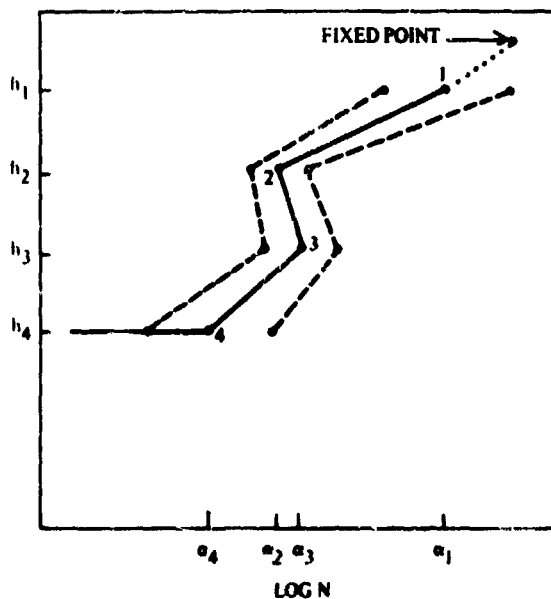


Figure 1.— Electron density model for $n = 4$.

The measure of fit of any electron density distribution to the data is taken to be the sums of the squares of the deviations of the computed values r_i from the data values R_i , where normalization is made by dividing through by the standard errors:

$$s = \frac{1}{m} \sum_{i=1}^m \left(\frac{r_i - R_i}{\sigma_{R_i}} \right)^2$$

Figure 1 shows the model used. The electron density distribution is specified by parameters α_j at a set of heights h_j in the profile. Logarithmic interpolation is made between the points, so that there are two things to find here: first, the distribution of heights, which is determined by the type of data available; and second, given those heights, we need the values of the parameters α_j that are the logs of the electron densities at the heights h_j .

We consider the problems in reverse order, first assuming that we know the distribution of heights and treating the problem of finding the parameters α_j , and later considering the distribution of heights.

THE DERIVATIVES, $\partial R/\partial \alpha_j$

For both problems we need derivatives $\partial r_i/\partial \alpha_j$ denoting the rates of change of reflection coefficients with changes in the model parameters. The solution for $\partial R/\partial \alpha_j$, hence for $\partial r_i/\partial \alpha_j$, is the problem we have recently solved for rapid computation.

The full-wave solution for R is an integral

$$R = \int \frac{dR}{dh} dh$$

which is carried out by Runge-Kutta integration. The expression for dR/dh is given by Budden [195] as

$$\frac{dR}{dh} = \frac{-ik}{2} \left\{ S^{(21)} + S^{(22)} R - R S^{(11)} - R S^{(12)} R \right\}$$

The initial condition for R is given by an analytic computation [Sheddy, 1968b]. The $S^{(ij)}$ are 2×2 complex matrices that are functions of the incidence angle and susceptibility matrix, which in turn is a function of the normalized plasma, gyro, and collision frequencies, X , Y , and Z , respectively, and direction of the magnetic field.

The normalized electron density X depends in turn on the model parameters α_j . We want to find derivatives of R with respect to α_j , which is the derivative on an integral:

$$\frac{\partial R}{\partial \alpha_j} = \frac{\partial}{\partial \alpha_j} \int \frac{dR}{dh} dh$$

We now take the derivative inside the integral:

$$\frac{\partial R}{\partial \alpha_j} = \int \frac{\partial}{\partial \alpha_j} \frac{dR}{dh} dh$$

At first glance this does not appear to save us anything because we still have to do the integration. However, we know that each parameter of the model affects only a certain region of the profile above which the derivative is zero, so that the integral is zero. Within the region affected by the parameter we integrate an expression obtained by differentiating dR/dh with respect to the parameter α_j :

$$\begin{aligned} \frac{\partial(dR/dh)}{\partial \alpha_j} = & \frac{-ik}{2} \left\{ \left[\frac{\partial S^{(21)}}{\partial \alpha_j} + \frac{\partial S^{(22)}}{\partial \alpha_j} R - R \frac{\partial S^{(11)}}{\partial \alpha_j} - R \frac{\partial S^{(12)}}{\partial \alpha_j} R \right] \right. \\ & \left. + \left[S^{(22)} \frac{\partial R}{\partial \alpha_j} - \frac{\partial R}{\partial \alpha_j} S^{(11)} - \frac{\partial R}{\partial \alpha_j} S^{(12)} R - R S^{(12)} \frac{\partial R}{\partial \alpha_j} \right] \right\} \end{aligned}$$

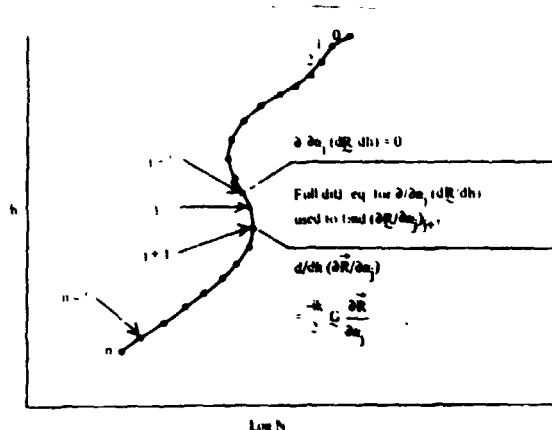


Figure 2 - Three regions of profile pertinent to obtaining j th derivative, $\partial R/\partial \alpha_j$.

In the region below the slabs affected by α_j this equation is linear and homogeneous since the first term in square brackets is zero. This is what saves us, because it turns out that integrations for linear homogeneous differential equations can be performed without knowledge of initial conditions. You can insert the initial conditions at the end.

Figure 2 indicates these three regions. In integrating we do four things. We perform the integration for R . We integrate through the pairs of slabs affected by the parameters α_j and store the value of each integral $(\partial R/\partial \alpha_j)_{j+1}$ at the bottom of the pair of slabs. We also integrate the linear, homogeneous set of equations through each slab and store the value of the integral A_j at the bottom of each slab, where A_j is such that

$$\left(\frac{\partial R}{\partial \alpha_j} \right)_{j+1} = A_j \left(\frac{\partial R}{\partial \alpha_j} \right)_j$$

and R is a vector with the same four elements as the matrix R . We can then work our way back up the profile, forming the product of A_j 's and computing

$$\frac{\partial R}{\partial \alpha_j} = A_{n-1} A_{n-2} \cdots A_{j+1} \left(\frac{\partial R}{\partial \alpha_j} \right)_{j+1}$$

So the time to get the whole set of derivatives does not really depend on how finely the profile is divided. We get the whole set of derivatives in about the same time, whether the profile is divided into 10 slabs or 100 or whatever.

SEARCH TECHNIQUE

Once we have the derivatives we use a combination of standard search (function optimization) techniques. The measure of fit s is to be minimized. We know something about the first and second derivatives of s :

$$\frac{\partial s}{\partial \alpha_j} = \frac{2}{m} \sum_i \frac{1}{\sigma_{R_i}^2} (r_i - R_i) \frac{\partial r_i}{\partial \alpha_j} \quad \frac{\partial^2 s}{\partial \alpha_j \partial \alpha_j'} \approx \frac{2}{m} \sum_i \frac{1}{\sigma_{R_i}^2} \cdot \frac{\partial r_i}{\partial \alpha_j'} \cdot \frac{\partial r_i}{\partial \alpha_j}$$

What we know forms an idealized parabolic surface of n dimensions (fig. 3), and along that surface we make essentially a Fletcher-Reeves search, which defines a quasi-steepest descent path down the surface.

The minimum of the real function s is then searched for along this path by a Fibonacci one-dimensional search, which involves about half a dozen full-wave solutions for each sounder frequency and accounts for most of the computation time.

Once the minimum is found along this path we have completed an iteration. For a rough approach to the minimum about half a dozen iterations may be sufficient.

RESOLUTION IN THE PROFILE

Now consider the problem of choosing the heights in the profile. The development of algorithm for choosing such heights was begun fairly recently. The philosophy here is that we do not want heights h_j in the profile for which electron density cannot be determined from the data. We can't have too many in, but on the other hand we want all that are warranted by the data.

The basis for choosing the heights h_j is the mapping of the uncertainties in the data parameters into an error band on the profile. Such a mapping is described in detail by *Shellman* [1970], and the mathematics are simplified in this presentation. The width of the error band on the profile is determined both by the uncertainties in the data and by the covariance between profile parameters α_j . The width of the error band at the height h_j is

$$\sigma_{\alpha_j} = \sqrt{I_{jj}}$$

where I_{jj} is the j th diagonal element of the inverse of the variance-covariance matrix. The latter has elements

$$\sum_i \frac{\partial r_i}{\partial \alpha_j} \frac{\partial r_i}{\partial \alpha_{j'}}$$

The rather simple algorithm for choosing the h_j — determining the resolution — is then the following: If at three successive points in the profile the error band is wider than a chosen upper threshold value, the middle point is deleted; and if the error band is narrower than a chosen lower threshold value, points are added on each side of the point. The threshold values are not critical.

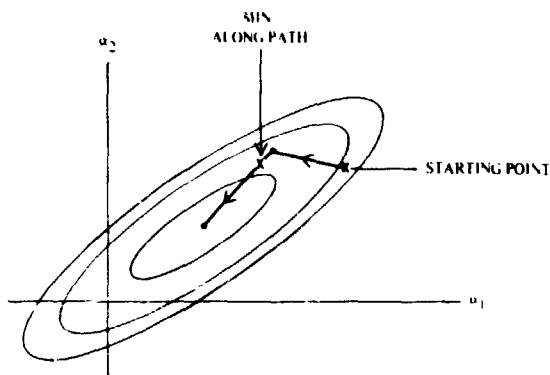


Figure 3. — Idealized parabolic surface, $s(\alpha_1, \dots, \alpha_n)$, defined by derivatives $\partial R / \partial \alpha_j$, $j = 1, \dots, n$. In the figure, $n = 2$.

AN EXAMPLE

Figure 4 shows the iteration steps in one particular case. In this example the variance-covariance matrix was normalized by dividing each j th row and j th column by $[\sum (\partial r_i / \partial \alpha_j)^2]^{1/2}$; hence, the matrix was actually a "correlation matrix." The principle is the same, however.

We start out with three (open circled) points in the profile. The algorithm for choosing heights then added a point at about 80 km, and one iteration in the numerical search led to the profile with the long dashed line. The second iteration led to the profile with the short dashes.

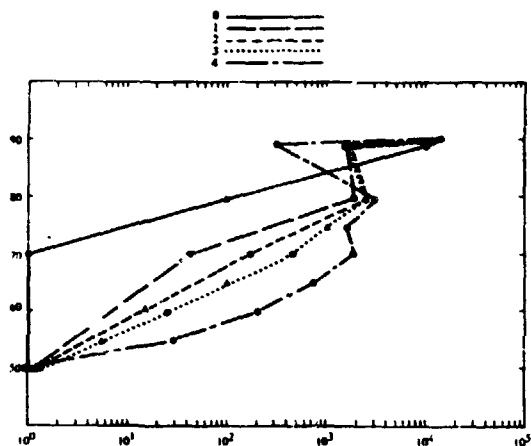


Figure 4. - First four iterations in search for profile which fits data taken 23 March 1971, 1300 MST. Data set consisted of $|R_p|$, $|R_s|$ and $|R_{p1}|$ at 8.3, 13.2, 21.7, 27.5, and 33.5 kHz. Phase data, ϕ_p , ϕ_s and ϕ_{p1} was added in subsequent iterations.

At that point the algorithm added a point at 60 km. Then we go to the dotted line in the next iteration and three points are added, indicated by the triangles. The next iteration then took us to the chain line. We carried out three or four more iterations, but the final profile was approximately the form of the last one shown.

In general, the algorithm for choosing the heights of the profile leads to about the same number of heights as we have data parameters, which is encouraging because that is what we would expect; that is, we get the same amount of information out that we are putting in.

It should also be mentioned that the collision-frequency profile has been treated in the same manner as the electron-density profile. That is, it has been assumed that the collision-frequency profile is also an unknown to be found from the sounder data.

REFERENCES

1. Budden, K. G.: The Numerical Solution of Differential Equations Governing Reflection of Long Radio Waves from the ionosphere. *Proc. Roy. Soc., London, Ser. A*, 227, pp. 516-537, 1955.
2. Paulson, M. R.; and Theisen, J. F.: NEL Ionospheric Sounding System. *J. Bur Ships*, vol. 11, pp. 3-6, 1962.
3. Sneddy, C. H.: A Curve-Fitting Technique for Obtaining Electron Density Profiles Consistent with VLF-LF Reflection Data (Abstract). *International Scientific Radio Union*, 1968 Spring URSI Meeting, p. 50, 9-12 April, 1968a (Program), National Academy of Sciences.
4. Shetty, C. H.: A General Analytic Solution for Reflection from a Sharply Bounded Anisotropic Ionosphere. *Radio Sci.*, vol. 3(8), pp. 792-795, 1968b.
5. Shellman, C. H.: Determination of Error Limits on Electron-Density Distributions Derived from VLF/LF Sounder Data. NELC Research Report 1676, 1969.
6. Shellman, C. H.: Electron-Density Distributions in the Lower Ionosphere with Associated Error Limits Derived from VLF and LF Sounder Data. *Radio Sci.*, vol. 5, pp. 1127-1135, 1970.

THE JOST-KOHN INVERSION PROCEDURE

Reese T. Prosser

Dartmouth College

N73-11627

With reference to my discussion of the Jost-Kohn procedure in Chapter 6, there are some conditions that must be imposed on the situation before reasonable results can be expected. These conditions are considered briefly here.

This discussion is restricted to the quantum-mechanical problem, but now in three space-dimensions, without assuming any radial or other symmetry of the potential. We consider the problem

$$\nabla^2 \phi + \omega^2 \phi = V \phi$$

In Chapter 6, I suggested that we search for a solution of the form

$$\phi(x, k) = e^{ik \cdot x} + \int G(x, x'; \omega) V(x') \phi(x') dx'$$

where V is the potential, G the associated Green function, k the incoming momentum, and k' the outgoing momentum, with $|k| = |k'| = \omega$. I said that in the far-field limit this solution would tend to something like

$$\phi(x, k) = e^{ik \cdot x} + T(k', k) \frac{e^{i|k||x|}}{4\pi|x|} + o\left(\frac{1}{|x|}\right)$$

that is, as an incoming plane wave plus an outgoing spherical wave weighted by a "reactance matrix" $T(k', k)$ of the form

$$T(k', k) = \int e^{-ik' \cdot y} V(y) \phi(y, k) dy$$

Earlier I implied, but didn't make explicit, that the scattering data of the problem are all contained in this reactance matrix, this far-field behavior being uniformly good for large $|x|$, and the correction terms being $o(1/|x|)$.

The Jost-Kohn procedure assumes that we have some information about the reactance matrix, from which we want to obtain the original potential V . The situation is better understood if we embed the problem in a family of nonphysical, purely mathematical problems, defined by *nonlocal* potentials. Specifically, we let this problem be a member of a larger class of problems in which the potential has the form

$$V \phi = \int V(x, x') \phi(x') dx'$$

We begin by counting dimensions. For a potential of this form, we must find a function of six variables. We thus need a six-parameter family of data, and that is supplied by the complete reactance matrix.

Proceeding formally, we express the reactance matrix T in terms of the potential matrix V . Since $\psi = \psi_0 + GV\psi$ and $T(k', k) = \langle \psi'_0 | T | \psi_0 \rangle = \langle \psi'_0 | V | \psi \rangle$, where $\psi_0 = e^{ik \cdot x}$ and $\psi = \psi(x, k)$, we have

$$\langle \psi'_0 | T | \psi_0 \rangle = \langle \psi'_0 | V | \psi_0 \rangle + \langle \psi'_0 | VGT | \psi \rangle$$

or, in operator form,

$$T = V + VGT$$

Thus

$$\begin{aligned} T &= (I - VG)^{-1} V \\ &= V + VGV + VGVGV + \dots \end{aligned}$$

This argument gives us a formal series, the Born series, for the matrix T in terms of the matrix V . The Born series converges, in general, only if the potential is sufficiently small. Specifically, we must introduce a suitable space of operators A, B, \dots acting on square-integrable functions, and then define a suitable norm for these operators such that $\|AGB\| \leq \|A\| \|B\|$, where G is the appropriate Green's function, and require that, in this norm,

$$\|V\| < 1$$

Clearly, the Jost-Kohn inversion procedure cannot be expected to yield anything useful unless the Born series converges, because the entire procedure is based on this series. We are therefore restricted immediately to weak potentials without bound states. The same argument applies to the Jost-Kohn procedure for other types of scattering problems, where it also depends on the convergence of a perturbation series.

The interesting thing, however, is that this requirement, seen here to be necessary, is also essentially sufficient. Suppose I insist that the norm of V is not only less than one, but also less than one-half:

$$\|V\| < \frac{1}{2}$$

Then I see that, since

$$T = V + VGV + VGVGV + \dots$$

we must have

$$\|T\| \leq \|V\| + \|V\|^2 + \|V\|^3 + \dots$$

or

$$\|T\| < \frac{1}{2} + \frac{1}{4} + \frac{1}{8} + \dots = 1$$

Thus, we can invert the equation

$$T = V + VGT$$

for V in terms of T :

$$\begin{aligned} V &= T(I + GT)^{-1} \\ &= T - TGT + TGTGT - \dots + \end{aligned}$$

giving V as the sum of a formal series that converges because $\|T\| < 1$.

This is all the Jost-Kohn procedure does. It is a very simple-minded procedure, which is not at all sensitive to any of the details of the problem. It is sensitive only to the fact that the problem must be solvable by convergent perturbation series.

This isn't the whole story, of course, because in practice we are looking for local potentials, and we are given only part of the complete reactance matrix. We generally know the scattering data only for certain aspect and scattering angles. For example, we know the case where k' is $-k$ —that is, the backscattering data for all aspects. Or more commonly in quantum mechanics, we know the scattering data for fixed aspect and all scattering angles. In either case, we know some but not all of the scattering data. Counting dimensions again, we see that a local potential is a function of three variables, so that in general we will need a three-parameter family of information, and that's just what we have in the two cases mentioned above.

Now the question is: If we know some part of the scattering data—some three-parameter part—and we know that the potential is local, can we then obtain the remaining scattering data, by analytic continuation or some other method, to construct the complete reactance matrix? This is where the papers of Faddeev mentioned by Calogero (Chap. 5) are so important; Faddeev apparently does provide just exactly that kind of information.

At this point, we can see that we have established a one-to-one correspondence between a certain class of weak potentials, local and nonlocal, and a corresponding class of weak reactance matrices. Moreover, this correspondence is described by convergent perturbation series in both directions. Now let's suppose that the original potential was in fact local. The inversion procedure can be carried out without this extra knowledge provided that the complete reactance matrix is known. That the potential is local is irrelevant to the inversion procedure, but is vital in the preliminary process of obtaining the complete reactance matrix from a suitable partial description. It would be highly desirable, I think, to find computationally effective ways of performing this preliminary process.

Suppose I look at the first Born approximation only, taking the series for the T matrix and ignoring all but the first-order term

$$T = V + \dots$$

If I invert this equation I find simply that in this approximation $T \cong V$. If we write this equation in the momentum space representation, and assume the potential is local, we obtain

$$T(\mathbf{k}', \mathbf{k}) \cong V(\mathbf{k}' - \mathbf{k})$$

If we know, for example, the backscattering data at all aspect angles, then we know $T(\mathbf{k}', \mathbf{k})$ for $\mathbf{k}' = -\mathbf{k}$, and so

$$b(\mathbf{k}) = T(-\mathbf{k}, \mathbf{k}) \cong V(-2\mathbf{k})$$

or

$$V(\mathbf{k}) \cong b\left(-\frac{\mathbf{k}}{2}\right)$$

In other words, in this approximation the momentum space description of the potential is essentially given by the backscattering data. This is Bojarski's result (discussed earlier in this chapter), translated into the quantum-mechanical problem. His very pretty result is, I think, capable of extension to all of these scattering problems.

One can say something qualitative now about how good the above approximation is. It is particularly good at high energies, as is well known, because the Born approximation becomes better as you go higher in energy. This observation is the analog of the one that Bojarski's result is a physical optics approximation valid for high frequencies, and the higher the better.

There are many ramifications of the Jost-Kohn procedure that are not yet clear. However, the subject can probably be pursued with profit in various directions to provide both qualitative information and effective calculation procedures for all these different scattering problems.

DISCUSSION

Moses: I was going to suggest that this procedure be tried for the one-dimensional case from plus to minus infinity, where things are much simpler. In fact some years ago I applied it to the one-dimensional problem, x going from minus infinity to plus infinity, and did the first two iterations. I wanted to compare the first two iterations with the first two iterations of the Gelfand-Levitan one-dimensional equation. Much to my surprise they were the same. So I conjecture, without proof, that term by term the Jost-Kohn procedure agrees with the Gelfand-Levitan iteration procedure if they converge. The Gelfand-Levitan iteration procedure seemed to converge if there were no bound states.

Newton: You are talking about the Marchenko version of it?

Moses: No, no. I am talking about the one-dimensional, Kay version, x going from plus to minus infinity.

Chadan: I think this approach which you explained has one drawback. As you say, the first term is good at higher energies. That means that it is good for the inner part of the potential in r space. Now, at high energies we know that the potential approach to many of our problems is not good. The potential approach is good for low energies, and there you have to know the outer part of the potential.

Prosser: I agree completely.

Moses: It is just that this is simpler to treat, I think, for electromagnetic theory and perhaps for the model.

Chadan: Yes, because if the frequency is large, then you see lots of details of the scatterer.

Calogero: I think the fact that if you expand the Gelfand-Levitan equation or the Marchenko equation or this approach, they will always give the same result, term by term. But this is not surprising. This could not be otherwise because we know that so long as we stay within the class of nonsingular potentials the scattering amplitude is holomorphic at small values of the coupling constant, and therefore you have the power expansion in the coupling constant, which is convergent in the neighborhood of $g = 0$ (let's call g the coupling constant), and therefore the two power functions cannot but coincide; the same is true for the inverse problem, so it is certainly so.

Moses: Normally one puts the coupling constant in front of the potential.

Calogero: Yes, but also the inverse is true, that if you put the coupling constant in front of the reflection coefficient and you expand the potential as a power series in this coupling constant, this is also convergent provided the reflection coefficient satisfies certain conditions.

Moses: That's right.

Calogero: But these are just the conditions that correspond to the requirement that the potential be nonsingular, and this has to do with the asymptotic behavior in k .

Moses: But this is an inverse problem. You want the condition on the reflection coefficient.

Calogero: These conditions are easy to find: They are just the conditions on the asymptotic behavior of large k . It has to decay at large k sufficiently rapidly so that when you iterate you don't get divergences.

Moses: That's all I wanted to say. Incidentally I have to say that when bound states were present, the iteration serves, for the Gelfand-Levitan equation did diverge. You have to remove the residues corresponding to bound states.

Sabatier: There is a recent series of papers by Barringer, and other nuclear physicists, on the off-shell continuation of t , and its use in inverse problems.

THE LOCATION OF THE POLES OF THE SCATTERING MATRIX

Peter D. Lax
New York University

and

Ralph S. Phillips
Stanford University

! N73-11628

This report describes a number of recent results bearing on the direct problem; that is, they relate geometric properties of the scattering obstacles to analytic properties of the associated scattering matrix.

Our results deal with exterior problems; for simplicity, let's take the acoustic case, governed by the wave equation

$$u_{tt} - \Delta u = 0 \quad (1)$$

in the exterior of an obstacle O , on which u satisfies the boundary condition

$$\frac{d}{dn} u = 0 \quad (2)$$

Other boundary conditions such as $u = 0$ and $du/dn + Ku = 0$ will also be considered.

We have studied *near-field* phenomena—the asymptotic behavior near the obstacle of solutions whose initial values are zero outside of a bounded set. What happens asymptotically to such a signal depends on the geometry of the obstacle. The key property seems to be whether there are reflected rays that stay inside the hull of the obstacle for arbitrary lengths of time. If there are such rays, the obstacle is called *capturing*. If on the other hand all rays are reflected out to infinity after a maximum sojourn time, the obstacle is called *noncapturing*. Convex obstacles are noncapturing par excellence, and so are star-shaped ones.

Our conjecture for noncapturing obstacles is that there exists a sequence of complex numbers μ_j , $j = 1, 2, \dots$,

$$0 > \operatorname{Re} \mu_1 > \operatorname{Re} \mu_2 > \dots \rightarrow -\infty \quad (3)$$

and functions $w_j(x)$, $j = 1, \dots$, depending only on the obstacle, such that every solution of the wave equation in the exterior whose initial values have compact support has an asymptotic expansion of the form

$$u(x, t) \simeq \sum a_j e^{i\mu_j t} w_j(x) \quad (4)$$

the numbers a_j depend on the solution in question. This conjecture has been proved for convex bodies by *Ludwig and Morawetz* [1969] and *Phillips* [1969]. For star-shaped bodies, *Lax et al.* [1963] have shown a weaker result:

$$u(x, t) = O(e^{-kt}) \quad (5)$$

where k is some positive constant that depends only on the obstacle.

The intuitive reason for the exponential decay indicated in Eqs. (4) and (5) is this: outside a noncapturing obstacle all rays go to infinity; according to geometrical acoustics signals follow rays except for a diffracted portion. This remaining portion again propagates out to infinity along rays, except for a diffracted portion, and so on ad infinitum. The signal remaining around the obstacle decays exponentially in t .

High-frequency signals suffer little diffraction; that is, they follow rays pretty closely even after a multiplicity of reflections. Therefore, around a capturing obstacle, which supports arbitrarily long reflected rays, there will be signals whose strength around the obstacle is hardly diminished no matter how large t is.

There is a bit of mystery about the numbers μ_j and functions w_j entering the expansion in Eq. (4); purely formally, $e^{\mu_j t} w_j(x)$ must be a solution of the wave equation so that

$$\mu_j^2 w_j = \Delta w_j \quad (6)$$

and w_j must satisfy the boundary condition

$$\frac{d}{dn} w_j = 0 \quad \text{on the boundary of } O \quad (7)$$

Clearly we are not dealing with ordinary, or even generalized, eigenfunctions that figure in the spectral resolution of Δ exterior to O , since the eigenvalue μ_j^2 is not real. To distinguish them from the eigenfunctions occurring in the spectral resolution of Δ , we shall call the w_j and μ_j occurring in Eq. (4) the *scattering eigenfunctions and frequencies* of Δ in the exterior of O , subject to the boundary condition in Eq. (7). Near infinity, the scattering eigenfunctions satisfy a radiation condition of the following sort:

For $|x|$ large, $w(x)$ can be expanded in a convergent series of the form

$$w(x) = e^{-\mu x} \sum_0 \frac{c_n(x)}{|x|^n} \quad (8)$$

where $c_n(x)$ are homogeneous functions of degree 0.

In theorem 5.5 of *Lax and Phillips* [1969], the following is proved:

Theorem 1 For any obstacle O there are infinitely many real scattering frequencies μ_j ; the number $N(\sigma)$ of μ_j between 0 and $-\sigma$ satisfies

$$c \leq \liminf \frac{N(\sigma)}{\sigma^2} \quad (9)$$

If O is star shaped, there is a corresponding upper bound for N :

$$\limsup \frac{N(\sigma)}{\sigma^2} < C \quad (10)$$

The quantities c and C are related to the geometry of the body by $c = 1.12r^2$ and $C = 1.12R^2$ where r and R are radii of inscribed and circumscribed spheres. In fact we conjecture that for any obstacle,

$$\frac{N(\sigma)}{\sigma^2}$$

tends to a limit as $\sigma \rightarrow \infty$.

A more recent result, contained in *Lax and Phillips* [1972], is the following:

Theorem 2 There are no scattering frequencies, real or complex, inside the circle

$$|\mu + \frac{1}{2R}| < \frac{1}{2R} \quad (11)$$

where R is the radius of a sphere containing O .

The following is a converse of theorem 1: If for large σ ,

$$\frac{N(\sigma)}{\sigma^2} \sim A$$

then the obstacle in question does not contain a sphere of radius $0.94(A^{-1/2})$, and is not contained inside a sphere of radius $0.94(A^{-1/2})$.

The following is a converse of theorem 2:

If μ is a scattering frequency contained inside the circle in Eq. (11), then the obstacle O is not contained inside a sphere of radius R .

If we could observe experimentally the location of the scattering frequencies, the last two results would shed some crude light on the overall size of the obstacle O . The chances of making such measurements are not very good, unfortunately. Since the natural characterization of these frequencies is in terms of the relaxation of the reflected near field, we need near field measurements, and these are not easy to come by. It is unlikely that we can measure accurately any of the scattering frequencies but the one that decays most slowly—that is, the one whose real part is largest. Thus, there is no chance of making any practical use of the converse of theorem 1, and only a slim chance that the converse of theorem 2 can be used to give a lower bound on the size of O .

DISCUSSION

Moses: There is another inverse problem which we haven't mentioned, and that is the synthesis problem. That is, you actually want to construct something that has a predetermined shape or something. What if you want to construct a hard body? You have to start off with something. A scattering operator has to have some requirements that indeed produce a hard body. These are necessary conditions for this.

In fact, in real electromagnetic engineering work this synthesis problem is a very important one. You actually do want bodies that have specific reflection. You do want transmission lines of that type. It is not a diagnostic. It is a construction.

Newton: I just wanted to make a brief remark about this work. This result, I believe, is quite different from what one could possibly hope to expect in quantum mechanics, because in the first place you don't have any signals with compact support in quantum mechanics, and partly because of that - of course, the cause is the same - there will be always asymptotic terms that go down as the inverse power of t . You don't have this kind of result.

Now, your second result there about the distribution of these, the leading term μ there, seems to me to be closely related to a result in quantum mechanics which, in the case in which the scatterer has finite support, relates the asymptotic decay, the relaxation length, which is exactly the analog of μ , to the size of the body. So I think that is quite directly related to it.

There is a reference to the quantum mechanical result in my book. In fact, it is discussed there.

Kay: How important is the requirement that the initial value have compact support? Suppose you have a point source and move it out to infinity so that it becomes a plane wave?

If you are treating this as an initial value problem, you start with a signal which is zero everywhere except in a bounded region; this would presumably satisfy the requirement. Then can you move that region out to infinity, increasing the strength as you go, so that you end up with an incident plane wave?

Lax: Well, the farther out you move it the longer you have to wait for it to come in.

What is important here is that we don't want things coming in after the initial signal arrives. We don't want rays continually coming in from infinity and interfering with this somewhat delicate decay behavior. That is why I insisted on a pulse.

Moses: I don't think it has to be. I mean, any pulse would be the scattering operator.

Kay: No, but the question here as related to the quantum mechanical application would be important, because there we would like to consider essentially a plane wave.

Moses: But you could use the time-dependent analog if you chose. The scattering problem would be the same.

Prosser: This depends effectively on the noncapturing conditions, because if you have a body that captures, then you can have arbitrarily slow decay rates.

Lax: That is right. If there is a phenomenon of capturing possible, as in this picture, then the following happens. Look at this interior cavity. If it is bitten off sufficiently sharply there is a reasonable way of closing it off as an interior problem, as an interior region, and this interior region will have eigenvalues λ_k , which correspond to vibration frequencies $i(\lambda_k^{-1/2})$.

Then the exterior problem will have μ_k , which will be approximately like these minus a small real part p_k , where p_k tends to 0 as k tends to infinity. For the higher and higher modes the exterior modes will be very close to the interior modes. These, however, are highly oscillatory modes, and the statement referred to the nonoscillating modes. The nonoscillating decay modes are related to the bulk properties of the object, and the highly oscillatory ones are to be detailed.

Shmoy: I think these results are very interesting, and the subject raises some interesting possibilities for inversion. I don't know how, but it seems that it should be possible. Certainly the sort of thing that some people in electromagnetics have been playing with—for example, it is possible to reconstruct the size of the sphere from the scattered field? Now nobody knows it is. It has been done in practice. This is another completely different approach to the same kind of question. Once you have done it for a sphere you can go to more complicated things, and perhaps from the knowledge of the radiative states you can deduce the properties of the system.

Calogero: Just one remark in connection with what you have just said. There is a tendency to think that once one has proved existence and uniqueness and perhaps a constructive procedure, the inverse problem is solved. That is not quite so, because you may have a very complicated construction procedure going through an integral equation, for instance.

Now, one can try to formulate other less ambitious inverse problems, in which you do not want to reconstruct the whole thing but you want to find some features of it, and then, of course, you may want to find interesting features, general shape or things like this. Then you may hope, if you are lucky, that you may be able to set up an inverse problem that will produce only these features, and this may be much simpler. The case I was talking about, the result obtained by Newton and Faddeev, is an example of this type, in which, in the context of the fixed angular momentum problem in quantum mechanics, the general problem is solved through the Gelfand-Levitan and Marchenko equations, but that involves solving an integral equation. It turns out that the explicit formulas can be obtained for the value of the potential, in fact all its derivatives, at the origin, which is, of course, much better, if one is interested in that feature of the potential, because that is an explicit formula.

Shmoy: This would give us a different approach than comes in from the other side. It appears to yield some information about the extent or moment.

Calogero: Yes.

Parker: I would just like to amplify your remark. This turns out to be exactly what happens in a specific geophysical inverse problem where, in the past, people have tried to be far too ambitious. The data, in fact, you can prove quite accurately, are completely ambiguous on the subject of the material [being sought from the measurement]. It has to do with the magnetization of sea mounts. They have been trying to construct the internal details of the magnetization.

But it turns out the ultimate purpose of their work was a bulk property, which could be found from the measurements, and so they have gone to all this great length of trying to parameterize the model, and limit the model in some way to make the problem unique, solve it, and then average the answer to find a bulk property. Now, in fact, if you can formulate the problem in such a way to go directly to the bulk property, everything is nice. I think this is a very important point you made.

Kay: I'd like to mention another point that occurred to me about the particular result, and that is that if you consider the long wavelength limit to scattering, that is, the body is small compared to the wavelength when you get Rayleigh scattering, the cross section is proportional to the square of the volume of the body, and there may be some relationship between that asymptotic result and the transient result that might be worth considering.

REFERENCES

1. Lax, P. D.; and Phillips, R. S.: *Scattering Theory*. Academic Press, N. Y., 1972.
2. Lax, P. D.; and Phillips, R. S.: Decaying Modes for the Wave Equation in the Exterior of an Obstacle. *CPAM*, vol. 22, p. 737-787, 1969.
3. Lax, P. D.; and Phillips, R. S.: On the Scattering Frequencies of the Laplace Operator in the Exterior of a Domain. *CPAM*, to appear 1972.
4. Lax, P. D.; Morawetz, C. S.; and Phillips, R. S.: Exponential Decay of Solutions of the Wave Equation in the Exterior of a Star-Shaped Obstacle. *CPAM*, vol. 16, pp. 477-486, 1963.
5. Ludwig, D.; and Morawetz, C. S.: The Generalized Huygens Principle for Reflecting Bodies. *CPAM*, vol. 22, pp. 189-205, 1969.
6. Phillips, R. S.: A remark on the preceding paper by C. S. Morawetz and D. Ludwig, *CPAM*, vol. 22, pp. 207-211, 1969.

MISCELLANEOUS COMMENTS

R. Parker: I would like to raise a question about an inversion problem that has not been considered here in detail. It is addressed to those people who are considering the electromagnetic scattering problem; it is a three-dimensional scattering problem that is routinely considered, and yet I don't know if any real mathematics has been done on it. It is the very important problem of constructing crystal-lattice electron densities and molecular structures from x-ray scattering.

The treatment is just a special case of very high-frequency electromagnetic scattering from a truly three-dimensional body with strong periodicity conditions imposed on the body. There is a definite element of nontriviality about it. The problem is formally very easily soluble if you have complex scattering information, but in fact one only has intensity measurements. The phases are lost to you by photographic film.

I was wondering if anyone had looked at the formal mathematics of this situation, to talk about the nonuniqueness, because people routinely claim they know exactly what the atoms, and so on, are.

J. Shmoys: We have a fairly active group in crystallography at Brooklyn Poly, and despite many attempts we have yet to establish contact between that group and our electromagnetics group, which is very much interested in periodic structures. There are semantic difficulties that seem to crop up.

I. Kay: The method of making holograms, which is to mix a known coherent field with the scattered field, will allow recording on film of phase information. It seems to me that something along those lines would be a more practical way of inverting this type of problem than considerations of intensity alone, which don't really have enough information to do the inversion. You need phase information.

R. Parker: But is that completely practical?

I. Kay: This isn't so farfetched. There are lasers in the high UV region right now. The reason it is impossible to invert pure intensity information is that it is incomplete. You can't reconstruct the phase information from the intensity. This is a well-known fact. You can do it the other way around.

J. Shmoys: We have heard the problem of identifying the phase function from the amplitude function in Professor Newton's talk.

I. Kay: It is the other way around. You can derive the amplitude from the phase, but not vice versa.

J. Shmoys: Holography is an interesting new area for exploration as an inverse method. I for one don't understand completely its possibilities and limitations.

K. Chadan: I have a comment on the existence of the kernel of the Gelfand-Levitan equation in quantum scattering theory. I will just discuss the case of the S-wave. We have the radial Schrödinger equation which reads

$$\phi'' + E\phi' = V\phi \quad (1)$$

and we are looking for the regular solution, which satisfies

$$\phi(0) = 0, \quad \phi'(0) = 1. \quad (2)$$

Now, on the basis of the Volterra integral equation for ϕ which incorporates the boundary conditions in Eqs. (2), or by using the well-known Poincaré theorem, you can prove that ϕ is an even entire function

of k for any value of r . You also find from the integral equation, that the asymptotic behavior of ϕ is given by

$$\phi \underset{k \rightarrow \infty}{\sim} \frac{\sin kr}{r} + O\left(\frac{e^{|\operatorname{Im} k| r}}{k}\right) \quad (3)$$

where the first term is just the free solution ϕ_0 . We see therefore that ϕ is an entire function of k of order 1 and type r for each value of r . Moreover $\phi - \phi_0$ is square integrable on the real k axis. It follows then from the Paley-Wiener theorem on the support of the Fourier transform of such functions that for each r ,

$$\phi - \phi_0 = \int_{-r}^{+r} K(r, t) e^{ikt} dt \quad (4)$$

where $K(r, t)$ is square-integrable with respect to t on $[-r, r]$. By integrating by parts, and using the fact that ϕ is even in k , one can put Eq. (4) in the familiar form given in Professor Newton's talk. So you see that here in the scattering problem the existence of the kernel $K(r, t)$ is just a consequence of Eqs. (1) and (2) and nothing else. This is in connection with Professor Moses' remark that the existence of the kernel may be related to the causality condition. This may be true in the time-dependent case with electromagnetic waves where you can make wave packets of finite size, and where you know that the speed of the particles is finite. In the non-relativistic case, particles may have any speed, and there are no wave packets with sharp fronts since only positive values of the wavenumber are available, so that there does not exist any simple way of formulating causality.

H. Moses: There is a time-dependent problem with causality, which Irv Kay brought up when he gave his talk, which has exactly the same scattering properties as this problem, and this is the one that I was referring to. It is a synthetic problem, but it has exactly the same scattering properties.

R. Phinney: I have some points and some questions which are not my own by any means, but represent things that I have culled from conversations in the past several days. To the seismologist the use of Gelfand-Levitan methods has a certain great attractiveness. Many of us are just learning about these methods, and, of course, we have been learning with realistic data, in fact acting upon this for some time.

The first question is addressed to the people in quantum mechanics. Are there experimentalists or teams with quantum-mechanical data upon which Gelfand-Levitan methods are being applied, or upon which they ought to be applied if they knew better?

R. Newton: As to the first part of the question, I think the answer is definitely no; they aren't being applied. Now, whether they ought to if they knew better is perhaps not answerable quite that easily. The trouble is that in most practical problems in which one really wants to know the force and where the force is not very well-known—say in a nuclear physics problem—the first kind of inversion problem is not really very practical because you don't know the data at the highest energies. The second kind of inversion problem, at fixed energy, would be much more practical if it were not for the fact that particles have spin, and that complicates matters considerably. In fact, the inversion problem at fixed energy for particles with spin, if the force depends on the spin, has not been completely solved. I had a graduate student who tried that as a project for his thesis. He didn't push it all the way through as an inversion problem, only part way, and it looks terribly complicated.

There are in addition spin-orbit forces, and Sabatier has written a paper on the inversion problem with fixed energy with the spin-orbit forces, and it is also very complicated. So to what extent, even after that problem is completely solved, it will be practical is quite open.

J. Shmoys: Let me just add a comment. In answer to this sort of question from a nuclear physicist who is an experimentalist, he felt it would be nice to try and apply these methods, even assuming the simple Schrödinger equation. I think it just has not occurred to many experimentalists that they ought to try to do something like that with their data. I am not sure what the reasons are, whether it is inertia or what, but I think it should be done, probably will be done eventually, but there is somehow no impetus at present.

R. Newton: Perhaps I should add something else. These techniques are sometimes useful not just for numerical inversions, but they are also useful for the construction of explicitly solvable examples. For example, the whole class of Bargmann potentials has been found by means of Gelfand-Levitan technique. Now, of course, once you have found those examples you don't need the Gelfand-Levitan technique any longer. They stand on their own feet. But that was the original way of finding them. So there you have a large class of potentials for which the Schrödinger equation is explicitly solvable. That is the class of potentials which lead to scattering matrices that are rational functions of k . Whenever the scattering matrix is a rational function of k , then you can write down the potential that leads to it explicitly for one angular momentum, that is for one L -value.

J. Shmoys: The equivalent sort of thing for transmission lines synthesis was Sharpe's theory.

H. Moses: I might mention that there is some interest in applying this to the ionosphere, provided it is simple enough, i.e., it is isotropic. There is an error analysis problem to which I think the geophysicists have contributed a great deal. You know, usually you don't have all the frequencies available. Also, in the past you haven't had amplitudes available, but now they are getting amplitudes and phases rather well for this scattering. It is just too recent. I think it is only in the last couple of years that material has come out on the ionosphere. There is an error analysis problem because you don't have all frequencies available.

R. Phinney: A slight extension of my question concerns the existence of limits to the amount of data, not necessarily the finiteness, but the cut-off in terms of frequency bands available. Certainly we would be very interested in methods which are both mathematically correct, in some sense, and practical, for applying to low and high frequency tails; to observed data which one might assume to be a continuum of perfectly good data in the middle of an accessible region. I take it from the remarks that this question has not yet become really important.

R. Newton: Oh, it is important, I think that is a very important question. In fact, that is a very interesting problem that hasn't been solved. I think that will be an extremely interesting problem to solve for many applications, namely, the problem of what class of potentials—if I describe it in quantum mechanical language—would fit certain phase shift curves known only in certain energy or frequency regions. That is a very interesting question, but that hasn't been solved.

H. Moses: It is related to the question of determining the character of the scattering operator corresponding to some knowledge of the potential. Suppose you know the potential cutoff in the nuclear physics case. Then you do know the asymptotic forms of the scattering operator, and perhaps something could be done, some sort of curve-fitting procedure. It is in its infancy.

J. Shmoys: I think that the situation in seismology may be worse, in a way, than in quantum mechanics or in plasma diagnostics, in that if you let the energies become large compared to the largest value of V , then you can say something readily about the asymptotic behavior, whereas if you have a nondispersive medium that tail is completely unpredictable. The dispersive character helps in quantum mechanics and in plasma diagnostics, whereas it would not in the nondispersive case.

R. Phinney: Since you raised the practicalities of the high frequency problem, in seismology there are two elements to it. The first is perhaps the simplest one, and that is the existence of finite dissipation which causes the broadening of spectral lines, and in fact, the inability to resolve spectral lines as free oscillations at short periods. This leads in practice to a complete traveling wave picture for the kinematics. This essentially removes high

frequencies. In fact, the high frequency signals that one observes have really fewer degrees of freedom in the data. Most of the signals have been completely attenuated and are not visible at all. This fortunately removes some of the infinity of degrees of freedom from the high frequency problem.

The other one is a little trickier, and this is related to the existence in seismology of multiple scales of the model, in which the scale length of the model can range from centimeters to thousands of kilometers. We have been using ray theory for some time, and are quite aware of improvements to ray theory, which enables one to look asymptotically at amplitudes, and yet, for instance, one can look at a sharp interface as a model for a particular physical situation, and study the reflection coefficients from this interface. Of course, they are frequency independent. One can then model this interface by some kind of ramp or Epstein potential which is smooth and get asymptotic behavior, and it is always analytic.

But, in fact, observed reflection coefficients don't behave this way at all. We get some practical fellows from the oil business who will say "Well, if we tune the band pass filters like this, we get really good signals. Otherwise you can forget the experiment." This means that the existence of the multiple scales is getting to us, and in a really practical sense the earth is not even a continuum, and we use continuum mechanics, of course, to derive these equations.

This leads to a remark about Tom Jordan's autocorrelation functions for the model. These are essentially smoothing windows which one decides the model ought to be seen through. They have been introduced in seismology, and yet there they are subject most to a priori assumptions for which you have no physical laws except, conceivably, for the earth's core, which is thought to be a liquid.

Yet in some of the other problems I have seen here, in plasmas and atmospheric problems, it is clear that the physics of the medium on small scales puts restrictions on how sharply the model can change, and so you can actually specify autocorrelations in the model, or at least inequalities to these things, and I think you can perhaps do something very useful this way. This, in a sense, provides a rational way of smoothing the data.

R. Newton: I would just like to say a few words about something that may be of use to some people. There was some discussion about the relation between solving the exact inverse problem and a discretization that one uses in approximations. I just want to give one little example of something which I think shows that the solution of an exact inverse problem can be very instructive, where one can be led astray by discretization.

Let me go back to the phase problem, in which a scattering cross section is given as a function of the angle and one wants to find the phase of the scattering amplitude. Suppose that an experimentalist is given a scattering cross section at a fixed energy as a function of the angle, and he finds that the curve that he gets is reasonably simple. Suppose he now fits that angular distribution by a Legendre expansion, in which he will of course use only a finite number of terms, preferably a small number of terms. He finds that to within his experimental error he can fit the cross section by a linear combination of Legendre polynomials up to order $2L$. In other words, he has now $2L + 1$ real parameters which he can use to fit his curve. Any experimentalist doing this would immediately draw the conclusion that if the cross section can be fit by a Legendre expansion up to order $2L$ then the amplitude must require only Legendre polynomials up to order L .

Now, the generalized optical theorem implies that each term in the Legendre expansion of the amplitude has only one real parameter, namely the phase shift. Here he has $2L + 1$ real parameters, and there he has only $L + 1$ real parameters to play with. Therefore he draws the conclusion that the generalized optical theorem implies a strong restriction on the possible angular dependence of a cross section.

But from the exact solution of the inverse phase problem, we know the following. The fixed-point theorem tells us that if you give me any reasonably smooth angular distribution for a cross section, if I now multiply that cross section by a parameter λ , and if I make λ small enough, then there always exists an amplitude that fits that cross section, no matter what the angular distribution is. Therefore, the generalized

optical theorem imposes absolutely no restrictions on the possible angular distributions that you can get. Thus the conclusion that was drawn here on the basis of the discretization was completely wrong.

Of course, if you assume that the amplitude has only a finite number of Legendre polynomials in it, then obviously it must terminate at L . If the cross section (the absolute value squared of the amplitude) terminates at $2L$, then the amplitude must terminate at L . But the point is that you may have to have infinitely many terms, and almost all amplitudes that fit a cross section that requires only Legendre polynomials, up to order $2L$, would require an expansion with infinitely many coefficients.

I think this shows that an exact solution of an inverse problem can teach you something about things that one otherwise might not be aware of.

9. WORKSHOP SUMMARY

P. C. Sabatier volunteered to author an overall summary to conclude both the workshop and these proceedings. The Editor felt that the scope of this task and the magnificent way in which it was accomplished warranted its publication in a separate chapter of this volume. J. Shmoys, session chairman, also has provided additional summary comments.

COMPARATIVE EVOLUTION OF THE INVERSE PROBLEMS

(Introduction to an Interdisciplinary Study of the Inverse Problems)

Pierre C. Sabatier

Département de Physique Mathématique*
Université des Sciences et Techniques de Montpellier (France)

ABSTRACT

1 N73-11629

The only way to extract maximum information from experimental results is to solve inverse problems. When a new study is undertaken, the first goal of the physicist is to obtain satisfactory fits. This can be done either by approximate methods or numerical computations. However, it soon becomes obvious that "good fits" can be obtained that correspond to fairly different data on the unknown quantities. The progressive realization of the consequences of this nonuniqueness imply an evolution of both the methods and the centers of interest in inverse problems. This evolution is schematically described in the present paper, together with the various mathematical methods used. A comparative description is given of inverse methods in scientific research, with examples taken from mathematics, quantum and classical physics, seismology, transport theory, radiative transfer, electromagnetic scattering, electrocardiology, etc. It is hoped that this paper will pave the way for an interdisciplinary study of inverse problems, leading eventually to the design of experiments that are most efficient to give information.

INTRODUCTION

Our knowledge of the world proceeds through "models." In any given model, the scientist defines a certain mapping M of a set C of functions into a set E . The "physical" functions are the elements of C that are mapped onto the set E_e of experimental results. The mapping M and the set—say C_e —of all the "physical" functions are altogether called the *physical law*.

In classical or quantum mechanics, the mapping consists of the "kinematics" and C is the set of the "dynamical" functions, such as the potential.

In seismology, the mapping is usually defined through the laws of propagation of elastic waves. The set C consists of several functions of position such as the density ρ , the bulk modulus K , the shear modulus μ , the local quality factor Q , which give a description of the earth. In geochemistry, C would have to include the concentration of various elements or compounds as a function of position and time. In geomagnetism, it would include electrical conductivity as a function of position.

In the diffraction of electromagnetic waves, C would include functions of position such as the conductivity of the diffracting body, and, most important, it would include the geometry of the body—namely, a boundary condition in the wave equations. The mapping is defined through Maxwell's equations.

The same general scheme holds in the transmission line scattering, but it is distinguished by its one-dimensional character and the fact that waves propagate along the line in two directions, left to right and right to left.

In studies of extended radiation sources, C includes the radial distribution of the emission coefficients (which in turn can give information on such parameters as temperature and density of the emitting matter). The mapping is defined through the laws of radiative transfer. More generally, in radiative transfer, as well as in neutron transport

*Physique mathématique et Théorique, Equipe de Recherche associée au CNRS. n°154.

theory, the set C is the set of properties of a semitransparent medium relating to the absorption, emission, and scattering of radiative fields.

In electrocardiology, C is the generator data, and the mapping is defined through the Poisson equation and the geometry of the human torso.

In the mathematical problem of "identification," M is defined by a partial differential equation, whereas C enters as either a boundary condition or as a parameter.

Throughout these physical or mathematical situations, we encounter the same general scheme. In the following, it is convenient to use the mechanical terminology and to call C the class of possible dynamics, C_e the class of effective dynamics, M the kinematics of the problem. Giving C and M , and defining the measurements, is giving a *model*. Now, let us assume that we *perfectly know the kinematics of a model*. There are two ways of using the model to increase our knowledge of the world. In the *direct problem*, the dynamical functions are known, either from general principles, or from previous information. The kinematics enables one to derive from the dynamical functions a set of numbers, which should fit the experimental data. In the *inverse problem*, the dynamical functions are derived from the experimental results and can be used in turn as fundamental information for further investigations.

The present paper is devoted to a comparative study of the inverse problems in physics. The most straightforward idea for such a comparison is to establish a classification of the physical situations in which an inverse problem has been of interest. Suppose first we put the examples quoted above in an order depending on the field concerned: classical and quantum mechanics, geophysics, electromagnetic theory, radiative transfer, electrocardiology, and mathematics. Suppose then we refined our classification—for instance, dividing geophysics into seismology, geochemistry, geomagnetism. Clearly the net result of such a classification would be to enhance the uniqueness of the inverse problems in various fields, and to make very difficult the classification of certain methods such as electrical sounding in geophysics.

A more persuasive way of making a classification is to describe the physical situations. In the example quoted above, we first encounter scattering studies, in which the experimental results give information on the asymptotic behavior of waves (quantum mechanics, seismology, electromagnetic theory). These studies reduce generally to Sturm-Liouville problems. There is, however, a large qualitative difference between the problems in which the main phenomenon is a stationary state of collision (quantum mechanics, electromagnetic theory) and the problems in which progressive waves are dominant (seismology). The second category of physical situations could be called the "transfer phenomena" in which the experimental results give information on the intensity of a transferred energy (radiative transfer, neutron transport). A third category would consist of "closed situations" in which the experimental measurements can be modified during experiments according to the information previously obtained (electrocardiology, electrical and some other soundings). It is clear that this method of classification is disappointing. The first category is fairly well defined and could justify a study by itself. The two other categories, although they are most important for applications, have many chances to contain several otherwise disconnected studies. For an extensive study of inverse problems, this classification may be of interest. For an introduction to interdisciplinary aspects, we will keep it as a secondary way of classification, not the main one.

A third way of making a classification would be to emphasize the mathematical distinction between a *well-posed* and an *improperly posed* problem. Since the definitions of these words exhibit important fluctuations, and are generally too precise for our purpose, let us say that we use them in the following sense: a well-posed problem is a problem in which the mapping M of C_e into E is a bijection, provided that E is the class of *all possible experimental results of a certain kind*. This means that for a well-posed problem, if a given element of E , a defined set of experimental results, is known *exactly*, then the dynamical function can be obtained *completely*. In practice, the experimental results are necessarily affected by errors, so that even in a "well-posed" problem the dynamical functions may be impossible to determine completely from experimental results, especially if the available inversion methods are not stable. The situation is much worse, however, in an improperly posed problem, in which E is either too small a class, so that an infinity of dynamical functions correspond to an "exact" set of "experimental results,"

or too large a class, so that certain elements of E do not correspond to any dynamics at all. For example, in quantum mechanics, if $\delta_o(E)$ is a known function *and* there is no bound state, the inverse problem for $l = 0$ may be well posed in our sense, although physically it is frivolous to speak of $\delta_o(E)$ for $E \rightarrow \infty$; if there is no restriction on the bound states, an infinity of potentials correspond to $\delta_o(E)$; and we describe the problem as improperly posed by *underdetermination*; if $\delta_q(E)$ is known for all l and all E , the conditions that have to be imposed on $\delta_q(E)$ to ensure the existence of a solution are unknown, and the problem generally is improperly posed by *overdetermination*. Obviously, improperly posed problems also may exist in which E is neither smaller nor larger than the class—say E_e —that is the image of C_e through the mapping M . The distinction between well-posed and improperly posed problems is of practical importance but it clearly cannot be the main motivation of a classification.

Ultimately, the most practical way of making a classification of the interdisciplinary aspects of inverse problems is to study separately the various steps through which the attempts of solution of such a problem may proceed. After each step, new questions are asked to increase the knowledge of the problem, and the following steps attempt to give them an answer. Since the questions are more and more refined, it is not surprising that the logical order of this "quest for information" is also, in many cases, the historical order of the studies of the problem. First came the approximate methods, followed by the trial-and-error methods. Then, generally, authors realize that the solutions are not unique, and ask a set of deep questions such as existence, uniqueness, constructibility, approximation theory, stability. Then the incomplete exact solutions come, in which subsets of E or C_o are used. The incidence of experimental errors and the statistical convergence of the systems identifications are studied apart. Ultimately, the complete solutions, the approximation theories, and, of course, the mathematical generalizations come to prove that the problem is now a "classical problem." Obviously, there is no real problem that has gone through the full cycle of evolution outlined here. Studies of problems concerned primarily with fundamental physics (e.g., quantum mechanics) went through a large part of the cycle, but did not pay much attention to the question of experimental errors. On the other hand, studies concerned with applied physics (e.g., the various soundings) paid most attention to the experimental errors and did not develop very interesting exact methods.

In this introduction to the interdisciplinary aspects of inverse problems, we outline the various steps of their evolution, with emphasis on the structures of the methods and the quantity of information referred to. Specialized methods of each discipline are minimized as a result. The similarities of the inverse problems in the various fields are emphasized while their differences are left to more detailed studies.

Examples are given of the problems noted earlier, particularly those in quantum mechanics and seismology, which exhibit almost all aspects of the cycle of evolution. Another reason for this emphasis is even better: the competence of the author rapidly decreases from quantum mechanics to electrocardiology!

MATHEMATICAL NOTATIONS

To summarize a physical model, we use the notation m with an index to denote the model, V with an index to denote a particular dynamics, and E with an index to denote a particular experimental result. A typical direct problem is therefore schematized as

$$(m_o): \quad V_o (V_o \in C) \rightarrow (M) \rightarrow E_o (E_o \in E) \quad (1)$$

and a typical inverse problem as

$$(m): \quad E_o (E_o \in E_e) \rightarrow V_i (V_i \in C_e) \quad (2)$$

In all the cases we study, it is possible to define C and E as complete metric spaces. The direct problem is formulated by

$$M(V) = E \quad V \in C \quad E \in E \quad (3)$$

Let E_e be a subset of E , which can be associated with the set of all possible experimental results of a certain type (for instance, all the functions $\delta_\ell(E)$ continuous and going to zero as E goes to ∞). The inverse problem is the problem of obtaining the solutions of Eq. (3) in C for all $E \in E_e$. The problem is well posed (in our sense), if there is one, and only one, solution for each element E of E_e . Notice also that our definition is essentially identical to the one introduced by Hadamard [1902] of a "problème parfaitement bien posé, je veux dire possible et déterminé." It is only more recently that a third condition was introduced to define a properly posed problem in Hadamard's sense, namely, that the solution of Eq. (3) depends continuously on E [Lavrentiev, 1967]. This condition is essential for stability.

APPROXIMATE METHODS

Classical Methods

In a number of wave propagation problems, the variations of the dynamical functions over a given wavelength are so small that diffraction and reflection can be neglected. Then wave groups propagate along classical trajectories to the next discontinuity surface; the parts of classical trajectories between two discontinuity surfaces are rays. These rays obey Fermat's principle: among possible paths joining any two assigned points in the medium, the actual ray is such that the travel time along it is extremal. If the radius of curvature of the surface discontinuities is sufficiently large, diffraction can also be neglected to study a discontinuity, which gives rise to a reflected and a transmitted wave, both propagating along rays, so that a completely classical model can be used in the medium.

Suppose now the problem has either spherical or cylindrical symmetry, and is a realization of one of the two following (equivalent) situations: the paths of propagation have turning points in the medium (quantum mechanics, seismology), or the source of radiation is an extended ball (or an extended cylinder). Then it is possible to relate measurable quantities to the dynamical functions by an Abel's integral equation.

Examples of such a result are shown in the inverse scattering problem at fixed energy in classical mechanics [Keller et al., 1956] and quantum mechanics [Firsov, 1953; Wheeler, 1955; Sabatier, 1966a; Vollmer, 1969; Miller, 1969], where the formula

$$\delta_\ell = \int_{\ell + \frac{1}{2}}^{\infty} \left[\lambda^2 - \left(\ell + \frac{1}{2} \right)^2 \right]^{-1/2} \lambda H(\lambda) d\lambda \quad (4)$$

obtained through the JWKB approximation, is typical. In Eq. (4), provided the potential allows only one "turning point," $H(\lambda)$ is bijectively related to the part of the potential $V(r)$ beyond the distance of closest approach r_0 ; δ_ℓ is the phase shift, as a function of the quantized angular momentum ℓ , and can be constructed, in certain conditions, from the experimental results.

A more useful example is the study of extended radiation sources where external measurements of the radiance produced by the source are used to deduce the radial distribution of the emission coefficient [Nestor and Olsen, 1960;

Freeman and Katz, 1960; Bockasten, 1961; Barr, 1962; Herlitz, 1963; Maldonado et al., 1965; Cremers and Birkeback, 1966; Minerbo and Levy, 1969]. Let us denote by $Y(y)$ the measurement of the radiance emitted by the extended radiation source along the edge of a circular slab of center O , in a cross section taken perpendicular to the axis of symmetry Oz (Ox and Oy being with Oz a rectangular triadron). Let $R(r)$ be the emission coefficient, which is assumed to vanish beyond a finite radius a . Then

$$Y(y) = 2 \int_y^a (r^2 - y^2)^{-1/2} R(r) r dr \quad (5)$$

A still more remarkable example of this type of result is shown in seismology, where the time T of travel along a ray can be determined, as a function of the ray parameter p , by

$$T(p) = \int_p^{\eta_0} \xi(\eta) (\eta^2 - p^2)^{-1/2} \eta d\eta \quad (6)$$

In Eq. (6), η_0 is equal to $r_0/v(r_0)$, where r_0 is the earth radius, $v(r)$ is the speed of propagation of the wave along the ray, at the point with radial coordinate r , and $\xi(\eta)$ is a function related bijectively to $v(r)$, provided that r/v decreases monotonely from η_0 to p as r decreases from r_0 to say, r_p ($r_p > r' > 0$). The problem is to recover $\xi(\eta)$ from $T(p)$ [Herglotz, 1907; Wiechert, 1907; Bateman, 1910; Slichter, 1932; Bullen, 1956].

Integral transforms similar to Eqs. (4), (5), and (6) also appear in the analysis of vertical incidence radio soundings of the ionosphere when the JWKB approximation is used. The relations between the group time delay of a pulse, or its transit time between two points, and a function directly related to the electron density are of this form, [Rydbeck, 1942; Kay, 1971], provided that the last function is monotonic, which ensures the uniqueness of the turning point. Again, a similar transform was given by Shioyama and Pirraglia [1971], in their studies of nonradially stratified media by trajectory methods. In diagnosing a two-dimensional refractive index distribution $n(r, \theta)$, expandable in Fourier series with coefficients $N_k(r)$, they prove that the Fourier component $I_k(b)$ of the phase shift of a ray, characterized by its impact parameter b and its angle α , is related to $N_k(r)$ by

$$I_k(b) = \int_0^{\pi/2} b N_k\left(\frac{b}{\sin \beta}\right) e^{ik\beta} (\sin \beta)^{-2} d\beta \quad (7)$$

which, for $k = 0$ and $k = 1$ reduces to a transform similar to Eqs. (4), (5), or (6). Other examples can be found in many cases where a classical approximation is used in a physical problem. The simplest one is probably the historical Abel's problem: determine the shape of a hill, given the time taken by a particle to go up and return as a function of the particle's initial velocity [Abel, 1826].

The formulas (4), (5), and (6) can be considered as integral equations for the dynamical functions [resp. $H(\lambda)$, $R(r)$, $\xi(\eta)$] when the "experimental functions" [resp. δg , $Y(y)$, $T(p)$] are known. The integral equation is of Abel's type, and can be solved exactly if $\delta(R + 1/2)$, $Y(y)$, $T(p)$, are known differentiable functions! For instance, the exact inverse of Eq. (5) is

$$R(r) = -\pi^{-1} \int_r^a (y^2 - r^2)^{-1/2} Y'(y) dy \quad (8)$$

Similar formulas hold in the other cases.

Analysis of the inverse problem— To use the above method, we first must assume that *the result to be found* makes the approximation possible. This is not a trivial assumption. In quantum mechanics, it means that we have to exclude the possibility of "orbiting." In seismology, it means that $v(r)$ has to increase steadily as r decreases. Now, it is possible to weaken this assumption—including, for example, special velocity variations [Bullen, 1950, 1963]—without implying any great mathematical difficulties, but only a small number of such complications can be taken into account. In any case, from the mathematical point of view, the class of functions in which the solution of the problem is desired is not very well defined.

A second difficulty comes from the experimental errors and fluctuations. It follows from them that the experimental function is *not known as a differentiable function*. Actually, the integral transform Eq. (6) is equivalent to half-order differentiation [Gorenflo and Kovetz, 1966], and therefore noise amplification is unavoidable. However, there has been no complete study of this instability up to now. Several authors have proposed numerical methods to solve the Abel's equation using Gaussian integration, interpolation formulas of low order [Nestor and Olsen, 1960; Frie, 1963; Bockasten, 1961; Edels et al., 1962], approximations of the data by smooth functions [Freeman and Katz, 1960; Cremers and Birkebak, 1966; Bullen, 1961, 1964], orthogonal function expansions [Herlitz, 1963; Maldonado et al., 1965; Gorenflo and Kovetz, 1966; Minerbo and Levy, 1969].

A third difficulty comes in quantum mechanics. In the other cases, if convenient assumptions are taken, the problem is well posed. In quantum mechanics, at the best, only integer ℓ values can be taken into account. The net result is that, even if the experimental errors were canceled, an interpolation would be necessary. It follows from Loeffel [1968] that the choice of the interpolation makes the problem well posed in the class \mathcal{U} of potentials whose $\int_0^\infty \rho |V(\rho)| d\rho$ is finite. Unfortunately, the choice of an interpolation consistent with JWKB approximation is difficult and in no case unique [Sabatier, 1966a]. The uniqueness appearing in Eq. (6) is misleading because the nonuniqueness appears in the step $\delta\ell \rightarrow \delta(\ell + 1/2)$.

Extensions and improvements of the results— In quantum mechanics and in seismology, as well as in the transmission line scattering, the formulas quoted above are obtained through the JWKB approximation. This approximation can be considered as the first term of a series of powers of the reflection coefficient [Bremmer, 1951] or similar parameters. Rather surprisingly, taking into account the next order of the approximation again yields an Abel's integral equation [Sabatier, 1966a; Vollmer, 1969]. The JWKB approximation and a first-order perturbation can also be combined; leading to an inversion method where both smooth and sudden (but small) variations of the dynamics are taken into account.

Other applications of classical methods to inverse problems— The ray method is commonly used for diagnosis in electromagnetic scattering, but its widest field of application is seismology. It has been used in almost all the determinations of earthquake parameters. Bullen [1963] provides a comprehensive review of these applications. More recent theoretical work is found in Gerver and Markushevich [1966, 1967]. It is interesting to note that the ray method led Mohorovicic, in 1909, to detect the famous discontinuity below the surface of the earth that has been named after him.

Usually, the most important experimental measurement used in the method is the arrival time of the wave although many other parameters, such as dispersion curves and amplitudes, can be taken into account. However, the filtering effect of the earth, together with the random distribution of the shells in the crust, prevent precise analyses of the spectra. On the other hand, under certain conditions, large deformations of the Mohorovicic

discontinuity can yield converging effects through which the amplitude measurements allow an easy diagnosis [Beaufils, et al., 1970].

Perturbation Methods

Let us assume that we know a certain exactly solvable model m_0 with which, for a given dynamic V_0 , depending on a few parameters, we can solve completely the direct problem

$$V_0 \rightarrow (M) \rightarrow E_0 \quad (9)$$

Suppose now V_0 is given a perturbation λV , and let ΔE be the modification of the experimental result. It may happen that ΔE can be expanded in powers of λ :

$$\Delta E = \lambda \Delta_0(V) + \lambda^2 \Delta_1(V) + \dots \quad (10)$$

where $\Delta_0(V)$, $\Delta_1(V)$, are certain functionals, which also depend on V_0 . A "perturbation method" is possible if $|\Delta E - \lambda \Delta_0 V|$ is much smaller than ΔE as ΔE goes to zero. Then, for reasonably small deviations to the solvable scheme in Eq. (9), it is possible to study the direct problem through replacing ΔE by $\lambda \Delta_0 V$. In addition, higher order of the perturbations expansions can be calculated in certain cases. The inverse problem can be handled if $\Delta_0 V$ is invertible. Once a value V_1 has been obtained from the measured ΔE , it is possible to use the expansion in Eq. (10) or an iteration process to improve the precision.

The perturbation method has been used in several inverse problems. In quantum mechanics, it yields the essential part of the methods of Fröberg [1947, 1948] and Hylleraas [1948] for inverse problem at fixed ξ ; it has also been used as a first approach toward the approximation problem [Sabatier, 1967a] in the inverse problem at fixed energy. In seismology, it has been studied by Knopoff [1960, 1961, 1962] in several type of problems [see also Mal and Knopoff, 1968, and the quoted references], including a study of higher orders of the perturbation expansion [Knopoff, 1962]. Applications to the inverse problems for surface waves have also been given by Jobert [1960, 1970], and to the inverse eigenvalue problems in seismology by Verreault [1965].

Let us now try to analyze more thoroughly the structure of perturbation methods in the inverse problems. (Such an analysis would have avoided well-known errors in some cases.) In typical problems, C and E can be given the structure of normed linear spaces, and M is a differentiable mapping of C into E , namely, a Fréchet differentiable functional. Then, if V_0 and V are elements of C , we can find a linear continuous functional $\delta M_0(V)$ such that

$$\Delta M = M(V_0 + V) - M(V_0) = \delta M_0(V) + r_0(V) \quad (11)$$

where $r_0(V)$ is a functional whose norm is infinitely small compared to $\|V\|$ as $\|V\|$ goes to zero:

$$(\forall \epsilon)(\epsilon > 0) \quad (\exists \delta)(\delta > 0)(\|V\| < \delta)(\|r_0(V)\| < \epsilon \|V\|) \quad (12)$$

where $\|-\|$ denotes the norm in C and $\|-\|$ the norm in E . Usually, C is a space of integrable functions and the linear function $\delta M_0(V)$ can be put in the integral form

$$[\delta M_O(V)]_\alpha = \int V(\rho) u(\alpha, \rho) d\rho = E_\alpha \quad (13)$$

where $u(\alpha, \rho)$ is a known function of α and ρ , and α is a parameter related to the structure of E . As an example, in quantum mechanics, at a given energy, in the (first-order) Born approximation

$$\delta_l = \frac{\pi}{2} \int_0^\infty V(\rho) J_{l+1/2}(\rho) J_{l+1/2}(\rho) \rho d\rho \quad (14)$$

In that case, α is a discrete parameter and E is a sequence space.

Let us now study the inverse problem in the first-order perturbation theory. Let us assume that the experimental function $E(\alpha)$ is perfectly known. Then Eq. (13) is a generalized moment problem, which in most inverse problems has at least a solution. Now let C_1 be a space of functions in which the g.m. problem has one solution only. Clearly, if the solution of Eq. (13) is sought in a set larger than C_1 , the problem is undetermined. This is the case for Eq. (14) if $V(\rho)$ is sought—say in $L_2(0, \infty)$. Actually, with the usual limitations on the δ_l , a unique solution could be obtained in the subspace of $L_2(0, \infty)$ containing only entire functions of order 1 and type 2 [Sabatier 1967a]. The conditions on $u(\alpha, \rho)$ for which Eq. (13) has a unique solution in a sufficiently large space of functions are generally not met in the physical inverse problems. The nonrealization of that point led Fröberg [1948] and Hylleraas [1948] to an incorrect guess [Bargmann, 1949; Levinson, 1949; Holmberg, 1952].

Let us denote by N_1 a linear functional that is an inverse of δM_O , with norm N_1 . Let us assume that $r_O(V)$ is Lipschitzian in a ball $(0, \delta'')$, and that the product L of N_1 by the Lipschitz constant is smaller than 1. (The first assumption is generally true in physical problem; the second one depends on N_1 .) Then the mapping $N_1[r_O(V)]$ is contracting in the ball of center O , radius $\delta' = \delta''(1 - L)$:

$$\|N_1[r_O(V_1)] - N_1[r_O(V_2)]\| \leq L \|V_1 - V_2\| \quad L < 1 \quad (15)$$

To obtain a solution of $\Delta M = E$, we now solve $V = N_1[E - r_O(V)]$ by iteration through the fixed-point construction procedure:

$$\begin{aligned} V^{(1)} &= N_1(E) \\ V^{(i+1)} &= N_1(E) - N_1[r_O(V^{(i)})] \end{aligned} \quad (16)$$

Clearly, the inequality in Eq. (15) and a well-known argument [see, for example, Shilov, 1965] enable one to guarantee that if $\|V^{(1)}\| < \delta'$, the sequence of Eq. (16) converges toward a unique solution V . Moreover, Eq. (14) is a constructive method: after a finite number of steps, it yields a result $V^{(i)}$ such that $\|V - V^{(i)}\|$ can be bounded, by $L^{i+1}(1 - L)^{-1}$. Therefore, to each inverse N_1 of the linear functional δM corresponds one inverse and one only of the functional M such that $\|N_1(E)\|$ is smaller than a certain quantity. Therefore, the differential δM yields a way to study the nonuniqueness of the inverse problem, but some other solutions may escape it. Notice, in particular, that, δM_O being linear, either N_1 is unique, or there exist an infinity of them, and, among them, some with a norm so large that the method above fails. It also yields a way to study the approximation problem and the fluctuations due to experimental errors, but again with the same limitation. From the theoretical point of view it is good for a first study but not sufficient for a complete one.

From the practical point of view, the use of a perturbation method has the same defects as classical methods. It requires a strong a priori assumption, and experimental errors are still a cumbersome complication.

Construction of perturbation methods— In most cases, the inverse problem is defined by a linear partial differential equation, and a certain set B of boundary conditions. Assume that the differential equation is of the form:

$$D_0(r)X = p(r)X \quad (17)$$

where $D_0(r)$ is a well-known differential operator in R_n , such that the solutions of

$$D_0(r)X_0 = 0 \quad (18)$$

with the boundary conditions B , are well known, and $p(r)$ contains the part of the dynamics which can be treated as a perturbation. Assume also that the Green's function corresponding to $D_0(r)$ and B is well known:

$$D_0(r)G(r, r') = \delta(r, r') \quad (19)$$

Then X is a solution of the integral equation:

$$X(r) = X_0(r) + \int G(r, \rho)p(\rho)X(\rho)d\rho \quad (20)$$

If the Neuman series for Eq. (20) converges, it yields a "perturbations expansion" for $X(r)$. The convergence of this series, however, is not always *necessary* to get a good approximation of $X(r)$ by the first-order (the Born approximation):

$$X_1(r) = X_0(r) + \int G(r, \rho)p(\rho)X_0(\rho)d\rho \quad (21)$$

Now, to make a measurement M_i of $X(r)$, we obtain the values of

$$\int M_i(r) \cdot X(r) dr \quad (22)$$

where $M_i(r)$ is defined from the experimental measurement, with the linear approximation i.e. Eq. (21) in place of $X(r)$; therefore, we get numbers of the form in Eq. (13).

NUMERICAL METHODS

The discussion of numerical methods logically should follow a complete analytic study of the mathematical problem. However, since the purpose of a physicist is to obtain results, numerical methods come in the study of inverse problems as soon as a constructive method is available, either exact, or approximate. One might question the relevance of numerical methods in a study of inverse problems. After all, most numerical methods are nothing but applications of the methods of solution of the direct problem to various dynamical functions, comparison with the experimental results, and choice of the best approximation. But the whole development of physics goes this way; we cannot encompass in our study the whole of physics! On the other hand, many analytic methods to construct a solution of inverse problems go also this way, since they proceed by successive approximations. In short, if we do

not include in our study all the fitting procedures, we have to make a distinction between the numerical computations according to the nature of the information they are able to extract from the experimental results, not according to the very structure of the method. Such a distinction is very arbitrary.

Throughout the papers devoted to numerical computations to fit the experimental results with the help of trial dynamical functions, we see two kinds of studies, through which, so to speak, the position of the guess is moving.

Completely Numerical Computations

These are methods in which the mathematical transcription of the physical processes is handled on the computer from the very beginning. For each step of the physical problem, there is a corresponding step of the program. The direct problem is solved in this way, for several possible dynamics, among which the program enables one to choose, so as to get the optimal fitting of the experimental results.

Let us take for example a one-dimensional process where a (partially) measurable quantity U is related to a dynamical function a of the medium, to be defined for $x \in [0, b_{N+1}]$, by

$$D(a, U) = 0 \quad (23)$$

and boundary conditions on certain components of U , say $U_k(0)$ and $U_k'(b_{N+1})$. The inverse problem is to infer the function $a(x)$ from measurements of a component of U at various points

$$U_0(x_i) = w_i \quad i = 1, 2, \dots, M \quad (24)$$

The problem can be discretized by introducing, to cover the medium, N homogeneous sections (b_i, b_{i+1}) , where the values a_i of the desired dynamical function are selected so as to minimize the sum of the square of the deviations:

$$S = \sum_{i=1}^M [U_0(x_i) - w_i]^2 \quad (25)$$

In dynamic programming the minimization is done at each step, through a function

$$f_K(c_1, c_2) = \sum_{i=1}^{M_K} [u(x_i) - w_i]^2 \quad (26)$$

The equation $D(a_i, U)$ is solved inside each section (b_i, b_{i+1}) , for $i = 1, 2, \dots, K$, with the boundary conditions c_1 and c_2 [which, ultimately, for $K = N$, should fit $U_k(0)$ and $U_k'(b_{N+1})$], and with appropriate continuity requirements of U at the boundaries of the sections. The minimization of f_K is done on the coefficients a_i at each step and M_K is the number of measurements on the first K intervals, with $K = 1, 2, \dots, N$. Sequences of functions are thus obtained, which solve the problem. The above way of proceeding is convenient for inverse

problems of transport theory [Bellman et al., 1967], diffusion processes, or radiative transfer, when the measurements are done all along the physical processes.

In most cases, however, the measurements are made only at a boundary of the medium. Then, if we call "trial results" the images of the trial dynamical functions in the mapping M , we see that the direct problem has to be solved completely for every trial result, and the deviation of these results from the observed ones has to be computed. The trial functions can be investigated in various ways.

In the "Monte Carlo" method, first a procedure is introduced to generate random dynamical functions. Next, it is necessary to specify a set of tests, whether direct or indirect, that must be satisfied by the functions to be chosen. Direct tests are a priori properties or conditions (e.g., bounds) on the functions. Indirect tests are essentially comparisons of the trial results and the experimental results. In these comparisons, the allowed differences—say δ_i —should be smaller than the errors (roundoff or discretization) coming from the numerical computations. In addition, the test is precise if the differences between the bounds defining the experimental data are smaller than the δ_i . Now, a random dynamical function belonging to the class defined by the direct tests is accepted if it satisfies all the indirect tests. Clearly, the great interest of this method is that it works as well for an underdetermined problem and for a well-posed problem. Even more important, this method is able to give nonprogrammed information: if the totality of the tests contain more information, about the dynamical functions and certain indirect tests than the direct tests and the experimental error bounds indicate, then at the end of an application of the method to a suitably large number of trial functions, the trial functions and/or the trial results corresponding to the nonprecise tests will lie in clearly discernible subregions in C . Such a method is particularly suitable for the (widely underdetermined) inverse problems of geophysics [Asbel, et al., 1966; Levshin et al., 1966; Keilis-Borok and Yanovskaya et al., 1967; Yanovskaya, 1963; Press, 1968; Anderssen and Seneta, 1970]. However, the method possesses also a great disadvantage, for the results of the trials already made are not used in the next trial. In the neighborhood of the best fit, if the "best fits" are isolated in C , a direct optimization method can save time, so that one ought to combine the two procedures. It is also possible to introduce an estimation procedure that formalizes the procedure for implementing and interpreting the results obtained [Anderssen and Seneta, 1970].

In a well-posed problem, a systematic method of optimization is generally preferable to a random method. Usually, the trial dynamical functions are chosen in classes described by a few parameters and the optimization is done on these parameters, a great variety of classical numerical methods being used to solve the direct problem of going from the trial functions to the trial results. See Hodgson [1963] for nuclear physics; Keilis-Borok and Yanovskaya [1967] for geophysics; Kagiwada and Kalaba, [1968], Bellman et al., [1963, 1965, 1967, 1968] for radiative transfer and diffusion processes; Gelernter and Swihart [1964] for electrocardiology; and Chavent [1970] for the general mathematical problem. Clearly, the advantage of choosing the trial dynamical functions inside a class depending on a few parameters only is that first the problem becomes well posed, and also much computer time can be saved. Unfortunately, the price to be paid is very high: a method constructed in this way can give only biased information, severely restricted by the class of trial functions. A good example is given by the numerical computations made in nuclear physics to fit the experimental cross sections by means of an optical potential. A three-parameter Woods-Saxon potential is generally used. Experimentalists are quick to note that even if all the possible data (differential cross section and polarization) are available, there are still some small ambiguities. They usually ascribed these small ambiguities to experimental uncertainties, and took for granted that changing the form of the trial functions would not make drastic modifications. However, for certain classes of trial functions, there can be drastic modifications [Sabatier and Quyen Van Phu, 1971].

Now we see that, in all these numerical methods, there is somewhere a guess on the dynamical functions. To apply any of them, and particularly in dynamic programming, it is necessary to assume that the problem can be discretized. This implies a "sufficient continuity" of the functions. Besides, in most cases, cheap parametrizations impose further restrictions. The method with the largest number of "degrees of freedom" would probably be the Monte Carlo method, if only an infinite computer time could be financially supported.

Reduced Numerical Computations

These are methods in which a preliminary step has reduced the problem to a simply defined mathematical one, usually an integral equation. In some cases, the preliminary step is done by approximate methods, hence introducing what can be called a badly defined, or a "hidden" guess. In many other cases, the preliminary step is exact. Then, the good definition of the second step makes it possible to study the uniqueness or the stability of the solutions much more fruitfully than in the completely numerical computations.

Examples of this type of approach are shown in radiation measurements, where a Fredholm equation of the first kind relates the unknown function $f(y)$ to the measured function $g(x)$

$$\int_a^b K(x, y) f(y) dy = g(x) \quad (27)$$

Such an equation has notoriously unpleasant mathematical properties. First, it can easily be seen that no solution exists, for a given $K(x, y)$, if $g(x)$ is an arbitrary continuous function. For example, one may check that if the function $x \rightarrow K(x, y)$ satisfies a linear differential equation, with coefficients depending on x only, $g(x)$ must also satisfy this equation. If $K(x, y)$ is a degenerate kernel, equal to $\sum A_r(x) B_r(y)$, no solution can be found unless $g(x)$ is a combination of the functions $A_r(x)$. Now, if the eigenfunctions of the kernel are known:

$$\int_a^b K(x, y) f_r(y) dy = \lambda_r f_r(x) \quad (28)$$

if no eigenvalue is zero, and if the eigenfunctions form a complete set in the space E , so that $g(x) = \sum b_r f_r(x)$, then $f(x) = \sum a_r f_r(x)$, where $a_r = b_r / \lambda_r$. This does not work for a degenerate kernel, which has at least one zero eigenvalue. Practically, these difficulties generate a lack of stability of the "approximate" solution with regards to the right-hand side of Eq. (27). There are several ways of reducing Eq. (27) to a matrix equation

$$K f = g \quad (29)$$

for instance, by numerical quadrature of the left-hand side of Eq. (27). But if the right-hand side is intended to be treated exactly, and if the number of points in the quadrature formula is increased, ill conditioning and instability will be the general lot. On the other hand, the problem of finding a smooth function $f(x)$ for which the remainder

$$e(x) = \int_a^b K(x, y) f(y) dy - g(x) \quad (30)$$

is in some sense small, and is well posed and physically meaningful. The smoothing can be done using some a priori knowledge of the unknown. Several methods have been used in this framework for inferring temperature profiles from stimulated emission measurements [Mazuel, 1968; Westwater, 1970]. In more recent studies, the minimization of $e(x)$ is understood as a minimization over a statistical ensemble. In the linear statistical method of Strand and Westwater [1968a, b, c], the coefficients of linear combinations are determined by minimizing the expected mean square error in the solution when averaged over a representative ensemble of profile and uncorrelated measurement errors.

Several examples of inverse convolution problems can be found in geophysics [seismology or magnetic interpretation, as in *Bott*, 1967] and in optics. The problems reducing to an Abel's integral equation, discussed earlier, are of this form. Other examples leading to the numerical inversion of integral operators are generally studied by well-known numerical methods. It is worth mentioning that in many cases, there are several possible reductions of a given physical problem (several integral equations giving a key to the solution). Some of these possible reductions may be better than others for numerical computations. It may also happen, in certain cases, that a new way of reducing the problem is physically better, because it automatically bars unphysical solutions. An example is given in ground electrical sounding, where the formulation used by *Kunetz and Rocroy* [1970] generates only sequences of positive resistivities.

Mathematical Foundations of the Numerical Methods in Inverse Problems

Rather recently, there has been interesting mathematical work on the general problems of convergence, optimization, statistical convergence, and filtering in well-posed inverse problems (there called "identification" and "estimation" problems). Several strong theorems have been proved by topological methods [*Bensoussan*, 1969].

Critical Comments on the Numerical Methods

The greatest asset of these methods is that they work in all the cases for which the direct problem can be solved numerically. Their greatest defect is that they give biased information. The trial functions have to be chosen in subclasses of C defined *essentially* for the purposes of the numerical calculations. Certain of the conditions are simultaneously physical conditions. There is no lack of information if all the conditions are really physical conditions—that is, conditions implied by the physical laws. In many cases, however, the so-called "physical" constraints follow from previous experimental tests. In other words, they follow from the known accumulation of previous "inverse problem" solutions in a certain subclass of C . Since these inverse problem solutions often are obtained by poorly conditioned trial and error methods, the constraints on this subclass of C should not be assumed in subsequent analysis. Rather, the investigation of the trial solutions should be as broad as possible, and the physicist should have always in mind that nothing proves that the path toward the complete knowledge of a physical law is unique.

On the other hand, if the problem is well posed, it reduces to a problem of "optimal control" and can be treated by well-known numerical methods [*Lions*, 1968].

FUNDAMENTAL QUESTIONS

To summarize the above analysis: What kind of knowledge do the approximate and the numerical methods afford? Given the experimental results, what is the dynamical function? These methods yield the answer, "Here is a dynamical function that for the given kinematics yields results approximately equal to the experimental measurements."

What kind of knowledge would we like to obtain? If we know, from previous information or from a general principle, a plausible dynamical function, then we would like to know whether the corresponding results do fit the actual measurements. If the only previously obtained information enables one to reduce the number of the arbitrary parameters characterizing the desired dynamical function, then it is interesting to use trial-and-error methods to solve the inverse problem. However, we must realize that the "proofs" obtained by these methods can always be questioned, while the counterexamples they may give cannot. Take, for example, the fitting of cross sections by means of a complex potential of the Woods-Saxon form. It has been recognized that a smooth variation of the potentials with respect to the kinetic energy is necessary for a good fit. We have every reason to believe that such a

dependence is "physically correct," since according to the theory the constructed potential should be equivalent with a nonlocal one. From the point of view of inverse problems, however, the fact that a static Woods-Saxon potential cannot fit the experimental results at all energies *does not prove* that *no* static potential can do it, since this inverse problem is improperly posed by underdetermination!

Therefore, what we would like to obtain from inverse problem studies is not only the most probable dynamical function, but the *total amount of information contained in experimental results*. This means that we must evaluate our results in terms of the following factors.

1. *Existence.* Does a dynamical function exist in C that gives, through the given kinematics, the proposed experimental result? In other words, if E_i is an "experimental result," is there an element C_i of C that M maps onto E_i ? When C_i does exist, it is called a *solution* of the inverse problem corresponding to the experimental result E_i .
2. *Uniqueness.* If E_i is given and C_i exists in C , is C_i unique? The answer is positive if the inverse problem is what we called "well-posed." From the mathematical point of view, a perfect experimental result is simply an element E_i of E . A real experimental result \bar{E}_i is actually the set of all the elements of E that are consistent with the measurements and their possible errors. Since in practical cases \bar{E}_i is given, not E_i , there cannot be uniqueness. Therefore, it should be understood that the question of uniqueness, as well as 3, 4, and 5 below, involve E_i , not \bar{E}_i . Only in the later step do we allow E_i to take on successively all values in the set \bar{E}_i .
3. *Constructibility.* It is not sufficient to know that a function exists. For a physicist, a much more important question is: Can a solution of the inverse problem be constructed for any element E_i of E ? By "constructive method," we mean one in which it is possible to obtain, after a convenient finite number of steps, any desired approximation of the solution.
4. *Complete constructibility.* When the answer to 2 is negative we also are led to the question: Can all the solutions of the inverse problem belonging to a given class C_1 ($C_1 \subset C$) be constructed? In the following, we call equivalent two solutions of the inverse problems corresponding to the same result E_i . Now, suppose we can construct a set ξ_1 of equivalent potentials. It is important to compare the elements of ξ_1 , or the elements of subsets of ξ_1 defined by some a priori "physical assumptions."
5. *Approximation.* We are therefore led to the question: Is it possible to define a distance in C_1 , and to get definite bounds for the diameters of the sets $\xi_1 \cap C_1$? Clearly, questions 3, 4, and 5 are essentially concerned with undetermined problems. Consider now a well-posed problem (if the problem is underdetermined, it is theoretically possible to restrict ξ_1 so as to obtain a well posed problem), and let us allow E_i to take various values in \bar{E}_i . It is most important to know whether ξ_1 show small or large perturbations when a small perturbation is imposed on E_i .
6. *Stability.* We now ask: With E and C_1 being metric spaces, is the mapping of C onto C_1 (the solution of the inverse problem) continuous? A related question concerns the efficiency of a series of tests: If a series of random tests yields results $\bar{E}_i^{(n)}$, and if the corresponding dynamical functions lie in the sets $\xi_1^{(n)}$, does the intersection of these sets steadily go "toward a unique element"?

We have touched on these questions in earlier sections. However, the studies described there were much more concerned with obtaining a solution, whatever it may be, than with information problems. In the next two sections, the quest for information is pursued methodically. The difference in the studies described does not reside (in most cases) in the nature of the methods they contain but in the quality of the information they seek.

FORMAL METHODS AND INCOMPLETE EXACT SOLUTIONS

Scarcely Parametrized Methods

The simplest way to construct an exact method is to reduce so drastically the class of possible dynamic functions that the direct problem can be exactly solved by analytical methods. The inverse problem then reduces to determining a very few parameters by elementary operations and matrix inversions. On the other hand, the necessary consistency conditions on the results are very strong. They can be fulfilled only if the accuracy of experimental results is very poor. In practice, this method is used mainly to give models that show the importance of a particular type of experimental result. More or less academic examples of this kind can be found in many areas, such as the one- or two-layer models of seismology [Ewing *et al.*, 1958].

Special Infinite Classes of Dynamical Functions

Physical considerations are sometimes sufficient to select in \mathcal{C} an infinite class of dynamical functions that, for some reason are physically more meaningful than the others. In quantum scattering, an example of this kind is shown in the class of superpositions of Yukawa potentials, which are favored by certain theoretical arguments:

$$V(r) = r^{-1} \int_{\mu}^{\infty} C(\alpha) \exp[-\alpha r] d\alpha \quad (31)$$

Another class corresponds principally to "simplicity" motivations, the class of truncated potentials—those that vanish beyond a finite radius. The same kind of motivations favors finite smooth bodies in electromagnetic scattering. In these scattering problems, the asymptotic properties of the waves can be represented with the help of a transition matrix, part of which is measurable in scattering experiments. From the properties of the special classes investigated, one can derive analytic properties of the transition matrix T , which in turn permit derivation of the matrix from its measurable portion by a process of analytic continuation.

In the class of Yukawa potentials for the quantum inverse problem at fixed ℓ [Martin, 1961; Chadan, 1962] and at fixed E [Martin and Targonski, 1961], the construction of the dynamical function from the transition function involves the use of a dispersion formula and a special iteration procedure. In the class of truncated potentials [Marchenko, 1957; Loeffel, 1968], this derivation is more formal. In electromagnetic scattering [Weston and Boerner, 1967], the field is readily calculated outside the convex envelope of the body. It remains to locate the surface of the (convex) body by taking into account necessary conditions on the boundary.

What is the value of these methods for the inverse problem? The inverse procedures are broadly formal. Instability for random perturbations of the data appears in the analytic continuations. In addition, the necessary consistency conditions on the experimental results are not really known (but they are probably not very strong). Therefore, these methods have hardly much practical value. On the other hand, they are important from the theoretical point of view, because their solution is generally unique and physically meaningful.

Although it is difficult to say that physical considerations have suggested the choice of separable potentials, we put here the study of their inverse problem because it has common characteristics with the above ones. Again, integral equations containing Hilbert transforms that come from dispersion relations are a key to the problem (the so-called Mushkelishvili equations). The inverse problem treated in this way is probably a good approach to the general inverse problem of nonlocal potentials. It has first been studied for theoretical reasons [Courdin and Martin, 1957, 1958; Chadan, 1958, 1967] but has more recently gained the favor of nuclear physicists [Bolsterli and Mackenzie, 1965; Mills and Reading, 1969; Tabakin, 1969], who could have saved much time by reading the previous references. One of the most remarkable successes of this problem is the result of Chadan [1967] concerning the possibility of

cancelling by a separable potential the phase shifts generated by a local potential, obtaining in this way a potential that is *transparent at all energies* and therefore cannot be distinguished from the potential 0 by scattering experiments! No example, we think, better demonstrates the surprising results that the physicist can reach in a study of question 2.

Formal Methods

The inverse scattering problem at fixed ℓ in quantum mechanics has given rise to a large number of studies from which some methods have emerged that have a large domain of applicability. The *Jost and Kohn* [1962] method, for instance, in which the transition matrix is expanded in powers of the interaction, has been used as well in electromagnetic scattering [*Prosser*, 1971]. The method of *Hylleraas* [1948], although it yields only one solution of an underdetermined problem, has recently gained some favor among chemical physicists. However, no method has been so generally applicable or given rise, by analogy, to so many other methods, as the Gelfand-Levitan method.

There are two aspects to the Gelfand-Levitan method. The first is its relation with the spectral problem for differential operators; the second is the "algebraic" structure of the method itself. The first one expresses the mathematical contents of the method, and the second expresses the constructive nature of the method for the study of inverse problems. We briefly consider the first aspect in the next two sections, referring the reader to the original paper of *Gelfand and Levitan* [1951] and the excellent review paper of *Faddeev* [1963]. As for the second aspect, it is interesting to draw the scheme of the method because it works for so many generalizations and analogous methods.

We are interested in the construction of the potential $V(r)$ from the phase shift of the regular solution of

$$[D(r) + k^2] \phi_V(k, r) = 0 \quad (32)$$

where

$$D(r) = D_0(r) - V(r) = \frac{d^2}{dr^2} - V(r) \quad (33)$$

and the normalization of $\phi_V(k, r)$ is chosen in such a way that $\phi_V(k, r) \sim r$ as r goes to zero. The asymptotic behavior of $\phi_V(k, r)$ is denoted by

$$\phi_V(k, r) = \frac{|f(k)|}{k} \sin [kr + \delta(k)] + o(1) \quad r \rightarrow \infty \quad (34)$$

The function $\delta(k)$ is the "experimental result." Now, the key of the Gelfand-Levitan equation is a function $K_0 V(r, r')$, referred to variously as the "transformation kernel" or the "generalized translation operator," which yields readily both the potential

$$V(r) = -2 \frac{d}{dr} K_0 V(r, r) \quad (35)$$

and the wave functions

$$\phi_V(k, r) = \phi_0(k, r) - \int_0^r K_0^V(r, \rho) \phi_0(k, \rho) d\rho \quad (36)$$

from the wave functions ϕ_0 corresponding to the potential zero (i.e. $k^{-1} \sin kr$). This kernel $K_0^V(r, \rho)$ is a solution of the partial differential equation

$$\left[D(r) - D_0(r') \right] K_0^V(r, r') = 0 \quad (37)$$

with the boundary conditions (35) and

$$K_0^V(r, 0) = K_0^V(0, r') = 0 \quad (38)$$

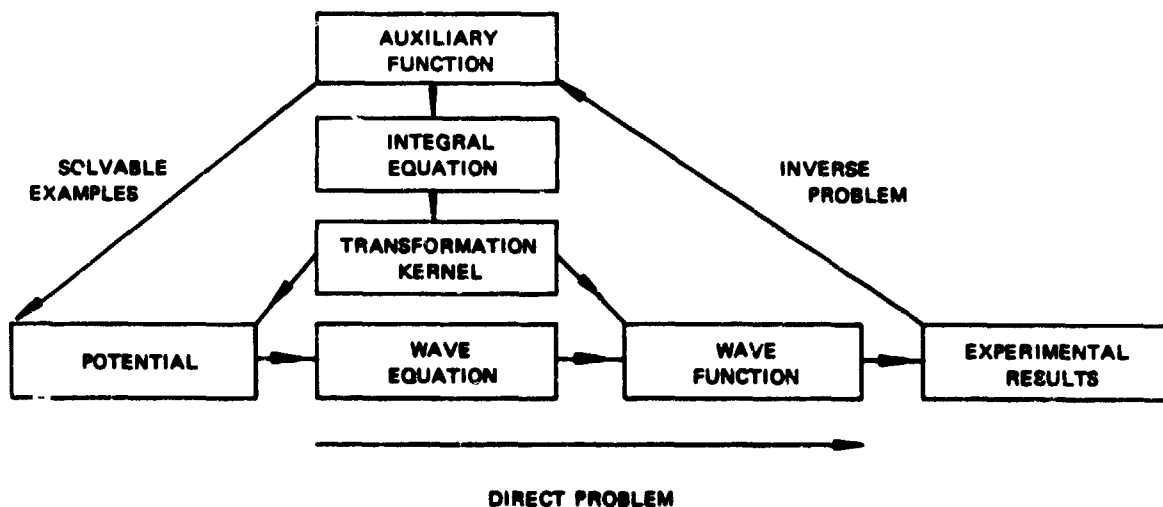
To construct $K_0^V(r, r')$, an auxiliary function $f_0^V(r, r')$ is introduced, which is a solution of the partial differential equation

$$\left[D_0(r) - D_0(r') \right] f_0^V(r, r') = 0 \quad f_0^V(r, 0) = f_0^V(0, r') = 0 \quad (39)$$

and which is related to $K_0^V(r, r')$ by the (Gelfand-Levitan) integral equation

$$K_0^V(r, r') = f_0^V(r, r') - \int_0^r K_0^V(r, \rho) f_0^V(\rho, r') d\rho \quad (40)$$

The logic of this procedure can conveniently be represented by a triangular diagram [Sabatier, 1968]:



Using an arbitrary solution of Eqs. (39) in place of the "auxiliary function" and constructing the transformation kernel yields examples of solutions of the Schrödinger Eq. (32), which can be exactly calculated if $f_0^V(r, r')$ is

conveniently chosen. As an example, if $f_0^V(r, r')$ is the product $\alpha \sinh(k_1 r) \sinh(k_1 r')$, we obtain a so-called "Bargmann potential" [Bargmann, 1949].

The second way of using the above scheme is to relate the experimental results to the "auxiliary function," obtaining a solution of the inverse problem. In the inverse quantum scattering problem at $\ell = 0$, where the above procedure exactly applies, this is done by relating $f_0^V(r, r')$ to $|f(k)|$ by a spectral expansion and relating $\delta(k)$ to $|f(k)|$ by a dispersion relation (discussed in a later section).

The logic of the machinery described above applies to a large number of inverse methods. Moreover, the analogies are generally not restricted to the general scheme but also imply analogous partial differential equations and integral equations. Let us sketch rapidly, by order of increasing complication, these extensions of the Gelfand-Levitan formalism.

Replacing $D_0(r)$ by $D_0(r) - \ell(\ell+1)r^{-2}$ readily yields the formalism for any ℓ . Replacing O by V_0 , the free-wave functions by the V_0 -perturbed wave functions, and V by $V - V_0$ readily yields the perturbed formalism. Replacing the wave functions by the Jost solutions and all the boundary conditions at the origin by asymptotic conditions at $+\infty$ we obtain the so-called "Marchenko formalism" [Marchenko, Agranovich and Marchenko, 1963; and the quoted references]. Both the Gelfand-Levitan and the Marchenko formalism can be extended by using vectors in place of the wave functions, yielding generalizations of the procedure to coupled equations [Newton and Jost, 1955; Krein, 1956], to the scattering of spin 1/2 particles by a tensor force [Newton, 1955; Agranovich and Marchenko, 1958], and to coupled channels [Cox, 1962].

To study the inverse scattering problem at fixed E , in which the "experimental result" is the set of the phase shifts δ_ℓ at the given energy, it is again possible to use the formalism quoted above, with the following replacements:

$$D_0(r) \rightarrow r^2 \left(\frac{\partial^2}{\partial r^2} + 1 \right) \quad (41)$$

$$\left. \begin{aligned} D(r) &\rightarrow D_0(r) - r^2 V(r) \\ k^2 &\rightarrow -\ell(\ell+1) \\ \phi_0(k, r) &\rightarrow \psi_0(\ell, r) = \left(\frac{\pi}{2} r \right)^{1/2} J_{\ell+1/2}(r) \end{aligned} \right\} \quad (42)$$

Moreover, the Gelfand-Levitan integral Eq. (40) and the generating formula in Eq. (36) are replaced by the Regge-Newton integral equation

$$K(r, r') = f(r, r') - \int_0^r K(r, \rho) f(\rho, r') \rho^{-2} d\rho \quad (43)$$

and the generating formula

$$\psi(\ell, r) = \psi_0(\ell, r) - \int_0^r K(r, \rho) \psi_0(\ell, \rho) \rho^{-2} d\rho \quad (44)$$

This formalism was introduced by *Regge* [1959]. To construct solutions of the inverse problem, *Newton* [1962] put a restriction on the class of possible dynamics by choosing a function $f(r, r')$ of the form

$$f(r, r') = \sum_{\ell=0}^{\infty} c_{\ell} \psi_{\ell}(r) \psi_{\ell}(r') \quad (45)$$

where $(\ell + 1/2)^{-1} c_{\ell}$ is absolutely and uniformly bounded. The c_{ℓ} can be obtained from conveniently bounded sets of phase shifts by matrix inversions [*Newton*, 1962; *Sabatier*, 1966b]. The potential thus obtained is not unique, unless sufficient decrease of the potential is imposed [*Sabatier*, 1966b]. Generalizations have been given, increasing the class of possible dynamics [*Sabatier*, 1966c, 1967a; *Coudray and Coz*, 1970].

Study of the inverse problem at fixed energy for spin-orbit potentials and for tensor forces requires a deep transformation of the partial differential equations and the integral equations, introducing a nonlinear coupling. A method for discovering procedures of this kind has been proposed by *Sabatier* [1968], who studied the spin-orbit case. *Hooshyari* [1970] gave a procedure to deal with both the spin orbit and the tensor potential. In both cases, the general logic of the method described is present.

Thus, generalizations of the Gelfand-Levitan machinery cover a number of inverse problems in quantum scattering theory. Still more important, many other inverse scattering problems can be reduced, often by elementary transformations, to mathematical problems formally identical to the inverse problems of quantum mechanics. Examples can be found in the inverse problem for strings [*Krein*, 1953; *Kac*, 1965], in the inverse problem for transmission lines [*Kay*, 1971, and quoted references], in the inverse seismic problem [*Blagoveshchenskii*, 1967; *Aki and Ware*, 1970], and in the inverse problem for plasmas [*Degasperis*, 1970].

Now, what kind of answer can be given to our fundamental questions by the procedures studied above. In certain cases, it is possible in this way to get a complete solution of an inverse problem. However, the completeness of the solution is not related to the structure of the method, but to the completeness of certain spectral expansions, so that its study is more conveniently deferred to the next section.

On the other hand, it follows readily from the very structure of the method that any "auxiliary function" that is a solution of a certain partial differential equation can be used to construct solutions of the direct problem. Therefore, to solve the inverse problem, it is sufficient to select a class of auxiliary functions having a one-to-one correspondence with the experimental results. Since this class is selected for technical reasons, it solves question 3; therefore, provided some attention is paid to convergence problems, it may solve question 1 and give a partial answer to 2. Furthermore, it may be relatively simple to study the continuity of the inversion procedure and to answer question 6. But, since the special class of dynamics in which a solution is sought is not chosen for physical but for technical reasons, there is no hope to answer questions 4 and 5.

The same conclusion is valid for all the methods available in the study of inverse problems in the relativistic case [*Corinaldesi*, 1954; *Verde*, 1958, 1959; *Pratt and Toll*, 1959; *Gasymov*, 1968; *Degasperis*, 1970; *Corbella*, 1970].

In concluding our discussion of the Gelfand-Levitan machinery and similar methods, we note that the theory of generalized translation operators [*Levitan*, 1964] is, in principle, the right tool for constructing new methods of this type [see, for example, *Coudray and Coz*, 1970], but in many cases more particular methods are more efficient [see, for example, *Sabatier*, 1968]. It also is worth noting that incomplete Gelfand-Levitan methods have sometimes been used for practical applications [see, for example, *Kim and Vasavada*, 1966, for elementary particles; and *Portinari*, 1967, for optics].

Problems Exactly Reducible to an Integral Equation

Some inverse problems can be exactly reduced to an integral equation. The fundamental questions 1 through 6 are then studied with regard to the integral equation. If this equation is linear, a great wealth of results is available in the literature, and all the questions can easily be studied. This does not mean that the studies are always pleasant. As we have seen, the Fredholm equation of the first kind, which is often encountered, necessitates several assumptions to answer question 1 and makes the study of question 5 difficult. However, a much more unpleasant case is the nonlinear equation encountered in the construction of the phase shifts from the cross sections [Newton, 1968; Martin, 1969]; in this problem, whose solutions were reviewed by Newton in Chapter 5, the only way of proving the existence of solutions is to use fixed-point theorems. The Banach theorem, which has the remarkable characteristic of giving a *constructive* method of solution, works when the cross section satisfies certain bounds, yielding one, and only one, solution. It is possible to increase slightly the domain of existence and uniqueness, and the domain in which the existence of the solution is guaranteed, by using the Leray-Schauder theorem. Unfortunately, even the largest domain is not very useful for applications.

Strong theorems of analysis have been used in certain inverse problems to give answers to questions 1 or 2 but are generally not very efficient in giving constructive methods [see, for example, Mireles, 1965; and Hunter and Bates, 1970].

Research of Condensed Properties

Usually, in inverse problems, the construction of the dynamical functions from the experimental results is quite complicated. Therefore, it is interesting to look for properties of the dynamical functions that can be readily derived from the experimental results. An example of these properties is given by the following exact relationship between the value of the potential at the origin, the scattering phase shift, known as a function of the momentum k , and the (negative) bound-state energies:

$$V(0) = -\frac{8}{\pi} \int_0^{\infty} \frac{d}{dk} [k\eta(k)] k dk - 4 \sum_n E_n \quad (46)$$

This formula was obtained by Newton [1956] through variational methods. Extensions of the result to the derivatives of the potential at the origin have been given by Faddeev [1957], Buslaev and Faddeev [1960], Percival [1962], Roberts [1964], Buskeev [1967], Calogero and Degasperis [1968], and in the relativistic case, Degasperis [1970]. These results were reviewed by Calogero in Chapter 5.

INFORMATIVE SOLUTIONS

As noted, the central problem in the previous section was the construction of the dynamics. Here the central problem is the determination of the amount of information contained in the experimental results, which can be pursued by three types of investigation. The first approach, the informative formulation of the problem, which applies to problems which are too intricate to be solved, consists of a direct appraisal of the spreading of the inverses of the functional M . The second consists of completely solving the problem in a priori well-defined mathematical classes, yielding what we call "a complete solution" of the inverse problem. It is then possible to perform the third way of investigation: that is, appraisal of the diameters of the subsets of equivalent potentials in C .

Informative Formulations

The best example of these formulations is the *Backus and Gilbert* [1968] method for the inverse problems in geophysics. It is assumed that the class C of all the possible dynamics (here, the earth models), is a linear-normed space (here, the continuous functions of r , with values in R_n), and one defines *each* way of measurement by a mapping M_i of C into E ; M_i is called a "gross earth functional." A finite collection of mappings M_i defines a set of acceptable dynamics. The authors assume that the functionals M_i are differentiable and replace them, locally, by their differential. This leads to the study of linear functionals M_i , which in a simple case have the form

$$g_i(m) = \int G_i(r) m(r) dr \quad (47)$$

where $G_i(r)$ is a known function of r . If the measured value of $g_i(m)$ is γ_i , the dynamical function $m(r)$ is defined by the generalized moment problem

$$\gamma_i = \int G_i(r) m(r) dr \quad (48)$$

Clearly, the problem has an exact solution if it is possible to construct a kernel $A(r_0, r)$ that is a linear combination of the $G_i(r)$ and such that, for any function $m(r)$ in the class of dynamics which are studied, and for any r_0 ,

$$m(r_0) = \int A(r_0, r) m(r) dr \quad (49)$$

If such a function is known, and is equal to

$$A(r_0, r) = \sum_i a_i(r_0) G_i(r) \quad (50)$$

the system of Eq. (48) has the solution

$$m(r_0) = \sum_i a_i(r_0) \gamma_i \quad (51)$$

The solutions to the inverse problems, except in very narrow classes of dynamical functions, are not unique, and it is not possible to construct a kernel $A(r_0, r)$ satisfying Eq. (49). But it is sometimes possible to construct approximate kernels $A(r_0, r)$, namely, unimodular linear combinations of the $G_i(r)$ that are in some sense nearly $\delta(r - r_0)$. More precisely, one selects, throughout the unimodular linear combinations of the $G_i(r)$, the one that is most nearly $\delta(r - r_0)$ (or, in Backus and Gilbert terminology, has the "best δ -ness"). Thus, such a method gives an "average" solution of the inverse problem; more important, it gives an appraisal of the deviation from each other of the equivalent solutions. This appraisal is readily related to the δ -ness of the kernel that can be effectively constructed from the knowledge of the functionals $G_i(r)$ and the experimental data. The deviation can be visualized in another way: one of its consequences is the existence of a dynamical function $m(r)$, which averages to zero over all but small length scales, and to the first order represents a possible deviation of any computed dynamical function. If this is the only source of the differences among the equivalent dynamics, there will be some smallest length such that all the acceptable dynamics have approximately the same average over an interval of length ℓ at radius r ; ℓ is called the resolving length at radius r . *Backus and Gilbert* [1970] were also able to determine how the errors in the data affect the resolving length of the data.

Now, the great value of this method clearly lies in its physically meaningful presentation of the consequences of nonuniqueness. In the frame-work of inverse problems in which the functionals are reasonably represented by their differential, it yields a fairly good answer to questions 4 and 5.

Earlier we showed how an "inverse" study completed on the differential of a mapping could be extended to the mapping itself, by using the Banach method of successive approximations. This extension is valid in a certain open ball of the normed space, with a radius related to the norm of the selected inverse mapping. Clearly, if we could find a covering of the whole normed space C by these open balls, the studies on the differentials would be fully extended. However, such a covering cannot be found. Where the distortion is too great, equivalent potentials escape this scheme, and only complete solutions (or, in lucky cases, random investigations) are able to detect them.

Complete Solutions

Let C be the set of all possible dynamical functions. A complete solution of the inverse problem would be a method of solution giving *all* the equivalent solutions in C . This definition is too strong, because it includes very unphysical complications, such as solutions that are almost everywhere equal and nevertheless should be considered as distinct solutions as defined above. Therefore, we use a slightly weaker definition in which C is replaced by "a mathematically *well-defined* subclass of C , which is dense in C for a reasonable norm and whose definition inside C is done *a priori* by simple mathematical requirements" (not *a posteriori*, to make the problem solvable, as in incomplete solutions). Let us discard the problems that reduce to a single equation and can be studied by well-known methods. The following complete methods are available in more complicated inverse problems.

In the inverse scattering problem at fixed ℓ , where the Gelfand-Levitan formalism applies, it is possible to show that $f(r, r')$ has the expansion

$$f(r, r') = \frac{2}{\pi} \int_0^\infty dk \phi_0(k, r) \phi_0(k, r') k^2 \left[|f(k)|^{-2} - 1 \right] + \sum_{j=1}^m C_j \sinh(k_j r) \sinh(k_j r') \quad (52)$$

where the k_j correspond to the bound states ($k_j^2 = -E_j$). This expansion *holds and is unique* for any given potential of a large class—say the class U —of continuous functions $V(x)$ such that

$$\int_\epsilon^\infty x |V(x)| dx$$

is finite for any $\epsilon > 0$. If the phase shift, as a function of k and the bound-states parameter, is consistently given, $|f(k)|$ and $f(r, r')$ can be constructed from which $K(r, r')$ and the potential can be derived. Therefore, the non-uniqueness in the inverse problem [where $\delta(k)$ only is known] comes from the bound-state arbitrariness.

In the Marchenko formalism, the relation between the auxiliary function $f(r, r')$ and the experimental results is still simpler than in the Gelfand-Levitan formalism; $f(r, r')$ is equal to a function of $(r + r')$, say $F(r + r')$, which has the expansion

$$F(x) = \frac{1}{2\pi} \int_{-\infty}^{+\infty} [S(k) - 1] e^{ikx} dk + \sum_j S_j e^{-k_j x} \quad (53)$$

where $S(k)$ is the (experimentally measurable) scattering function ($S(k) = \exp [2i\delta(k)]$), and the k_j and S_j refer to bound states. Again, this expansion exists and is unique for any potential in U . Clearly, the Marchenko method is simpler to apply to the inverse problem at fixed ℓ than Gelfand-Levitan's. This improvement is certainly related to the fact that the Marchenko approach reproduces more faithfully the physical path of the information. A further consequence is that the Marchenko method has applications in a larger area of scattering problems, such as the transmission-line inverse problems.

Each unique expansion of a function $f(r, r')$ has its origin in the spectral analysis of the boundary-value problem consisting of the (differential) wave equation and the boundary conditions required for its solution. If the dynamical functions belong to a suitably defined class, this boundary-value problem has, in the energy plane, a continuous spectrum coinciding with the entire positive real axis ($E > 0$), and, possibly, a finite number of non-positive eigenvalues. The corresponding eigenfunctions are a total subset in the space of functions $f(r, r')$ corresponding to the "well-behaved" class of potentials.

The spectral analysis has been applied by *Loeffel* [1968] to the inverse quantum scattering problem at fixed energy. He was able to prove that $f(r, r')$ has a unique spectral expansion, and that this expansion can be related to properties of the dynamical interpolation of the Jost function. Unfortunately, the Jost function is not related to the phase shifts in as direct a way as in the inverse problem at fixed ℓ , and no constructive method can be obtained in this way. However, this has given partial answers to questions 1 and 2.

The complete method given by *Sabatier* [1972] to the inverse scattering problem at fixed energy follows from a Fourier transform property of $f(r, r')$. It can be shown [Sabatier, 1970], for a subclass \tilde{E} of u that is not much narrower than U , that $f(r, r')$ can be written as

$$(rr')^{-1} f(r, r') = 4 \ell i m \int_0^\infty w^{-1} \sin w u F(u) du \quad (54)$$

where

$$w = [(r - r')^2 + 4rr'u^2]^{1/2} \quad (55)$$

$$F(u) = \begin{cases} F_1(u) & \text{for } u \leq 1 \\ F_2(u) & \text{for } u > 1 \end{cases} \quad (56)$$

where $uF_1(u)$ is a continuous function and $F_2(u)$ belongs to $L_2(1, \infty)$. The important point is that $F_2(u)$ can be chosen (almost) arbitrarily. Then the part of $f(r, r')$, corresponding to $F_1(u)$, say $f_f(r, r')$, can be expanded in a double Bessel series

$$f_f(r, r') = \sum_{\ell=0}^{\infty} c_\ell u_\ell(r) u_\ell(r') \quad (57)$$

where the c_ℓ are related to $F_1(u)$ and, most important, can be obtained by several matrix inversions from the δ_ℓ once $F_2(u)$ has been chosen. In that work, the author did not use the Regge-Newton formalism but an asymptotic form of this formalism, obtained by keeping only the asymptotic part of the operators [Sabatier, 1970]. Such a formalism has the advantage of avoiding as much as possible the dispersion of information in complicated functions of several variables. It enabled the author to put the problem in the form of a generalized moment problem for a

function bijectively related to the dynamical function, the so called "scattering structure" function. Somewhat analogous (but less useful) results can be obtained for the inverse problem at fixed ℓ using the Krein's method, and can probably be obtained in most scattering problems.

Approximation theory— Any complete method in which algorithms yield the equivalent solutions can theoretically be used to study the "approximation" question 4. However, from a practical point of view, an approximation theory is most easily constructed if it is possible to associate to the dynamical function a function of a well-known space, such as $L_2(0, \infty)$, in which approximation theories are well known. This is easy in the complete method given by *Sabatier* [1971], because of the properties of $F_2(u)$. It is possible to derive the diameter of the set of equivalent potentials characterized by given bounds on the derivatives, and to predict that for a static potential this diameter goes to zero as $E \rightarrow \infty$ and is generally smaller if the bounds on the derivatives are smaller. These results have been checked by numerical computations [*Sabatier and Quyen Van Phu*, 1971]. In the inverse problem at fixed ℓ , Krein's method has been used by *Melnikov* [1961] to study approximate methods.

Advantages of the complete methods— Clearly, the complete methods, when they are constructive, give answers to questions 1 through 5. The continuity of algorithms for the construction of the equivalent potentials can be checked to give answers to question 6. Is it possible to conclude that the complete methods are perfect? Unfortunately, they have two defects: First, they are very difficult to obtain, except in relatively simple cases; and second, they are complicated. However, these two properties may be the most constant characteristics of inverse problems.

MATHEMATICAL GENERALIZATIONS

The inverse methods involve many mathematical tools, most of which are beyond the average level of mathematical tools in physics. Therefore it is not surprising that an important body of mathematical literature on inverse problems has developed. Moreover, physical inverse problems have suggested the study of many mathematical inverse problems. After all, the problems of construction of a function of certain classes such as harmonic functions, or holomorphic functions, from its values on a boundary are in a certain sense inverse problems. If we limit our point to inverse problems readily originating from physics, it is worth mentioning the studies made to clarify or formalize numerical methods or improperly posed problems (see, for example, *Lavrentiev*, 1967). However, by far the most numerous mathematical studies connected with inverse problems have been those of inverse spectral problems (construction of an operator from a spectral function) and the theory of generalized translation operators. These subjects have been extensively studied by Russian mathematicians, in particular, Gelfand, Levitan, Marchenko, Krein, Agranovich, Faddeev, and Berezanskii, whose works have been cited many times (see also references to Naimark, Ljance, Gorbacuk, and Kac). The research of general properties of Sturm-Liouville problems, has been of interest in mathematics since the 1930's. Many general results [*Titchmarsh*, 1940] are of interest in the physical inverse problems. More particular results have been devised for the inverse problem in quantum mechanics [see, for example, *Borg*, 1947; *Levinson*, 1949a,b]. Other results could easily be used for physically interesting generalizations [*Crum*, 1955; *Friedman*, 1957].

All these mathematical studies have the outstanding characteristic of being rigorous: The class in which the solution is sought is correctly defined, integrals and series that are used converge, and except for some casual and trivial errors without important implications, the results can be taken for granted. Unfortunately, this is not always true in physical papers. Therefore, it is very important to enhance and increase the curiosity of mathematicians relative to the physical inverse problems. On the other hand, it also is true that mathematicians tend to study the simplest inverse problems with the weakest assumptions, instead of giving solutions of more difficult problems with stronger assumptions. Nevertheless, it is easy to see that among inverse problems, the ones which have awakened the interest of mathematicians are now the best known. Are they well known because they have been studied by mathematicians, or are they studied by mathematicians because they are easy to understand?

CONCLUSION

The direction of the evolution of inverse problems is determined by a qualitative evolution of the information sought. Physicists first try to obtain a solution of the problem, whatever it may be, either exact or approximate; for this purpose, approximate methods and numerical methods are the best tool. They soon discover that the dominant feature of the inverse problem is the lack of uniqueness of solutions, and it becomes necessary to determine whether this feature is due merely to experimental uncertainties (well-posed problem) or to the very structure of the problem (improperly posed problem). In any case, the problem moves toward the information questions. Roughly speaking, the physicist goes from

"Here are the experimental results. Give a dynamical function to fit them." to "Here are the experimental results. What amount of information do they contain?"

Apart from this general trend, the inverse problems exhibit many features in common, the origin of which can easily be ascertained. The Abel's integral equation is clearly a subproduct of a turning point in a classical trajectory. The perturbation method, which is a very general tool in functional analysis, produces, to the first order, linear functionals. The integral form of such a function being of a large generality in analysis, Fredholm integral equations come in a natural way. And the Gelfand-Levitan and derived methods can be related to properties of self-adjoint operators in scattering problems.

For an interdisciplinary working program on inverse problems, we think that these three tools—*classical methods*, *inversion of linear functionals*, and *spectral methods*—should be three chapters. A fourth one would deal with the *numerical methods of identification, with optimizing programs*. A fifth one, perhaps, would deal with statistical numerical methods. In all these studies, the six fundamental questions presented earlier must be studied and answered.

Now there is a further point which may be, in the future, of fundamental interest. In our opinion, inverse-problem studies should be used not only to explain experiments but also to plan them. For a given model, inverse methods can be used in designing experiments in a way that more information can be extracted from fewer experiments. When this is done, the studies of inverse problems will be the most important tool in scientific research.

ACKNOWLEDGMENTS

During the Fall 1970, the author was invited to give a course on "Inverse Problems in Physics" in the Indiana University Physics Department. This was certainly the best preparation to the present lecture. The author is very grateful to Professor R. G. Newton for giving him this opportunity, and for several interesting discussions. He is also indebted to the students attending that course for useful remarks.

DISCUSSION

Slomys: What kind of bias are we putting in when we restrict the class of potentials? It is not clear whether the Backus-Gilbert method, which appears to remove some of that bias or formalize the procedure, is the best. Whatever procedure is used, there is always the question of somehow smoothing the information. We either restrict the class of potentials which we are trying to construct or we smooth the data we obtain immediately and proceed analytically. This restriction must occur, and it is not obvious which end to put it in, at the data end or class of potential end.

Sabatier: In any case you have to put a priori constraints on the class of dynamical functions you are looking for. If you know what constraints you are making you are then OK.

Unidentified speaker: There is one aspect deserving more emphasis, which is illustrated by the Backus-Gilbert procedure. We should concentrate on how much information is contained in the data supplied by the physicist. These constitute measured values, including errors contained, in light of what we know a priori about dynamical variables. Then we should determine what this implies about the dynamical variables. The important question is, how much work do we have to perform to extract that information? That subject is in its infancy. The astonishing thing about Backus-Gilbert is that they get through with a feasible amount of computation.

Sabatier: I agree.

Parker: In the situation where you have your complete solution to the quantum mechanical scattering problem how do you describe the content of the physical measurements?

Sabatier: This deserves much study. Your question is the problem of stability. If the inverse mappings are continuous, and if they are differentiable, then you can associate to a ball in the space of experiment results, a ball in the space of dynamical functions. Then, if you take experimental results with only 15 points, you in fact know more than these 15 points. In the Gelfand-Levitan method you know 15 points. But from the a priori constraints you have on the potential, you can put inequalities on all the other values you don't know. In this case, then, you also have an infinite set. You have to put bounds on all the things you don't know and on the experimental errors.

Calogero: At what point one should smooth the results is still the question. One always smooths experimental data. One must do it sensibly stating errors.

Parry: In the matter of cost of producing data, and the aims set down in the whole field of geoscience, these measurements are extremely expensive. I would like to add another goal: We not only want the physical relationships and the information content, but also we would like to have a unique one-to-one correspondence between the physical parameters and the inverted parameter.

REFERENCES

- Abel: [1826] Résolution d'un Problème de Mécanique. *Journal Für Die Reine Und Angewandte Mathematik Herausgegeben Von Crelle*, Bd. I, Berlin.
- Agranovich, Z. S.; and Marcenko, V. A.: [1957] *Dokl. Akad. Nauk. SSSR*, vol. 113, pp. 951-954.
- Agranovich, Z. S.; and Marcenko, V. A.: [1958] *Dokl. Akad. Nauk SSSR*, vol. 118, p. 1055. [1963] *The Inverse Problem of Scattering Theory*, Gordon and Breach Pub.
- Anderssen, R. S.; and Seneta, E.: [1970] A Simple Statistical Estimation Procedure for Monte Carlo Inversion in Geophysics (Preprint).
- Asbel, I. Ja.; Keilis-Borok, V. I.; Yanovskaja, T. B.: [1966] A Technique of a Joint Interpretation of Travel - Time and Amplitude - Distance Curves in the Upper Mantle Studies. *Geophys. J. Roy. Astr. Soc.*, vol. 11, pp. 25-66.
- Backus, G. F.; and Gilbert, J. F.: [1968] The Resolving Power of Gross Earth Data. *Geophys. Journ. Roy. Astron. Soc.*, vol. 16, pp. 169-205. [1970] Uniqueness in the Inversion of Inaccurate Gross Earth Data. *Phil. Trans. Roy. Soc.*, vol. 266, pp. 123-192.
- Bargmann, V.: [1949a] Remarks on the Determination of a Central Field of Force from the Elastic Scattering Phase-Shifts. *Phys. Rev.*, vol. 75, pp. 301-303. [1949b] On the Connection Between Phase Shifts and Scattering Potential. *Rev. Mod. Phys.*, 21, p. 488.
- Barr, W. L.: [1962] Method for Computing the Radial Distribution of Emitters in a Cylindrical Source. *Journal Opt. Soc. Am.*, vol. 52, pp. 885-888.
- Bateman, H.: [1910] The solution of the integral equation which connects the velocity of propagation of an earthquake wave in the interior of the earth with the time which the disturbance takes to travel to different stations on the Earth's surface. *Phil. Mag.* (6), vol. 19, pp. 576-587.
- Beaufils, Y.; Mechler, P.M.; Rocard, Y.: [1970] Description générale des variations de réception sismique selon les régions de la France. Explication proposée. *Compt Rend. Acad. Sci. Paris*, t. 270, pp. 926-928, 1070-1072, 1122-1124.
- Bellman, R. E.; Kagiwada, H. H.; Kalaba, R. E.: [1963] An Inverse Problem in Dynamic Programming and Automatic Control. *J. Math. Anal. Appl.*, vol. 7, pp. 322-325. [1967] Dynamic Programming and an Inverse Problem in Transport Theory. *Computing*, vol. 2, pp. 5-16. [1965] On the Identification of Systems and the Unscrambling of Data I, II. *Proc. Nat. Acad. Sci. USA*, vol. 53, pp. 907-913. III *Ibid. J. Math. Anal. Appl.*, vol. 23, pp. 173-182 (1968).
- Bensoussan, A.: [1969] Sur l'identification et le filtrage des systèmes gouvernés par des équations aux dérivées partielles. *I.R.I.A.*
- Berezanskii, Ju. M.: The Uniqueness Theorem in the Inverse Problem of Spectral Analysis for the Schrödinger Equation. *Amer. Math. Soc. Transl.*, Ser. 2, vol. 35, pp. 167-235.
- Blagoveshchenskii, A. S.: [1967] *The Inverse Problem in the Theory of Seismic Wave Propagation Topics in Mathematical Physics*, vol. 1, Ed by M. Sh. Birman. Consultants Bureau New York.

- Bockasten, K.: [1961] Transformation of Observed Radiances into Radial Distribution of the Emission of a Plasma. *J. Opt. Soc. Amer.*, vol. 51, pp. 943-947.
- Bolsterli M.; Mackenzie, J.: [1965] *Physics*, vol. 2, p. 141.
- Borg, G.: [1946] *Acta Math.*, vol. 78, p. 1.
- Bott, M. H. P.: [1967] Solution of the linear inverse problem in magnetic interpretation with application to oceanic magnetic anomalies. *Geophys. J. R. Astr. Soc.*, vol. 13, pp. 313-323.
- Bremmer, H.: [1951] The W. K. B. Approximation as the First Term of Geometrical-Optical Series Symposium on the Theory of Electromagnetic Waves. Interscience New York.
- Bullen, K. E.: [1956] Seismic Wave Transmission in Handbook of Physics, *Geophysics*, Ed. I. S. Flugge, Springer-Berlag, Berlin, 1956.
- Bullen, K. E.: [1961] Seismic Ray Theory. *Geophys. J. Roy. Astr. Soc.*, vol. 4, pp. 93-105. [1964] Seismic Travel-Time Power Series Expansions Involving Terms in Δ^2 . *Geophys. J. Roy. Astr. Soc.*
- Bullen, K. E.: [1963] *An Introduction to the Theory of Seismology*, 3rd Edition, Cambridge University Press.
- Buslaev, V. S.: [1967] The Trace Formulas and Some Asymptotic Estimates of the Resolvent Kernel of the Three Dimensional Equation, *Topics in Mathematical Physics*, Ed. Pirman, Consultants Bureau, 1967.
- Buslaev, V. S.; and Faddeev, L. D.: [1960] *Dokl. Akad. Nauk. SSSR*, vol. 132, p. 13, Engl. Transl. *Soviet Math. Doklady*, vol. 1, p. 451.
- Calogero F.; and Degasperis, A.: [1968] Values of the Potential and its Derivatives at the Origin in Terms of the S-wave Phase Shift and Bound-State Parameters. *Journ. Math. Phys.*, vol. 9, p. 90.
- Chadan, K.: [1958] On the Connection Between the S-Matrix and a Class of Non-Local Interaction. *Nuov. Cim.*, vol. 10, pp. 892-910. [1967] On a Class of Completely Transparent Non-Local Two-Body Potentials. *Nuovo. Cim.*, vol. 47, pp. 510-525.
- Chadan, K.: [1962] *Nuovo Cimento*, vol. 24, p. 379.
- Chavent, Guy: [1970] Quelques méthodes applicables à l'identification de paramètres répartis dans les équations aux dérivées partielles. *I.R.I.A. Inf. 7008*.
- Coudray, C.; and Coz, M.: [1970] Generalized Translation Operators and the Construction of Potentials at Fixed Energy. *Ann. Phys.*, vol. 61, pp. 488-529. [1971] Construction of Relativistic Potentials at Fixed Energy. *J. Math. Phys.*
- Corbella, O. D.: [1970] Inverse Scattering Problem for Dirac Particles. Explicit Expressions for the Values of the Potentials and Their Derivatives at the Origin in Terms of the Scattering and Bound-States Data. *J. Math. Phys.*, vol. 11, pp. 1695-1713.
- Corinaldesi, E.: [1954] *Nuovo Cimento*, vol. 11, p. 468 et vol. 12, p. 469.
- Cremers, C. J.; and Birkebak, R. C.: [1966] Application of the Abel Integral Equation to Spectrographic Data. *Appl. Opt.*, vol. 5, pp. 1057-1064.

- Crum, M. M.: [1955] Associated Sturm-Liouville Systems. *Quart. J. Math.*, vol. 6, pp. 121-127.
- Degasperis, A.: [1970] On the inverse problem for the Klein Gordon S-Wave Equation. *J. Math. Phys.*, vol. 11, pp. 551-567. [1970] On the Inverse Problem for a Cylindrical Plasma. *J. Math. Phys.*, vol. 11, pp. 3392-3399.
- Edels, H.; Hearne, K.; and Young, A.: Numerical Solutions of the Abel Integral Equation. *J. Math. and Phys.*, vol. 41, pp. 62-75.
- Ewing, W. M.; Jardetzky, W. S.; and Press, F.: [1958] Elastic Waves in Layered Media. Mc Graw Hill New York.
- Faddeev, L. D.: [1957] *Dokl. Akad. Nauk. SSSR*, vol. 115, p. 878. [1963] The Inverse Problem in the Quantum Theory of Scattering. *J. Math. Phys.*, vol. 4, pp. 72-104.
- Firsov, Q. B.: [1953] *Zh. Exp. Teor. Fiz.*, vol. 24, p. 279.
- Freeman, M. P.; and Katz, S.: [1960] Determination of the Radial Distribution of Brightness in a Cylindrical Luminous Medium with Self-Absorption. *J. Opt. Soc. Amer.*, vol. 50, pp. 826-830.
- Friedman, A.: [1957] On the properties of a singular Sturm-Liouville equation determined by its spectral function. *Mich. Math. J.*, vol. 4, p. 137.
- Frie, W.: [1963] Zur Auswertung der Abelschen Integralgleichung. *Ann. Physik.*, vol. 10, pp. 332-339.
- Fröberg, C. E.: [1947] Calculation of the Interaction Between Two Particles from the Asymptotic Phase. *Phys. Rev.*, vol. 72, pp. 519-520. [1948] Calculation of the Potential from the Asymptotic Phase. *I. Ann. for Mat. Astr. O. Fys.*, B 34 A/28, pp. 1-16. [1949] *Ibid.* II. *Ibid.*, B 36 A/11, pp. 1-55.
- Gasymov, M. G.: [1968] An Inverse Problem of Scattering Theory for a System of Dirac Equations of Order $2n$. *Trudy Moskov. Mat. Obsc.*, vol. 19, pp. 41-112. [*Math. Rev.*, vol. 39, p. 2418].
- Gelernter, H. L.; and Swihart, J. C.: [1964] Numerical Computation of Body-Surface Potentials Produced by an Arbitrary Distribution of Generators in the Heart. *Electrophysiology of the Heart*. Pergamon 1964, p. 229-243. A Mathematical-Physical Model of the Genesis of the Electrocardiogram. *Biophys. Journ.*, vol. 4, pp. 286-300.
- Gelfand, I. M.; and Levitan, B. M.: [1951] On the Determination of a Differential Equation from its Spectral Function. *Amer. Math. Soc. Transl.*, Ser. 2, vol. 1, pp. 253-304.
- Gerver, M. L.; and Markushevich, V. M.: [1966] Determination of a Seismic Wave Velocity from the Travel-Time Curve. *Geophys. J. R. Astr. Soc.*, vol. 11, pp. 165-173 (1966). [1967] On the Characteristic Properties of Travel-Time Curves. *Geophys. J. R. Astr. Soc.*, vol. 13, pp. 241-246 (1967).
- Gorbacuk, M. L.: On the Spectral Functions of a Differential Equation of Second Order with Operator Coefficients. *American Math. Soc. Transl.*, Ser. 2, vol. 71, p. 177.
- Gorenflo, R.; and Kovetz, Y.: [1966] Solution of an Abel-Type Integral Equation in the Presence of Noise by Quadratic Programming. *Numer. Math.*, vol. 8, pp. 392-406.
- Gourdin, M.; and Martin, A.: [1957] Interaction Non-Locale Séparable et Matrices de Collision. *Nuov. Cim.*, vol. 6, pp. 757-779. [1958] Exact Determination of a Phenomenological Separable Interaction. *Nuov. Cim.*, vol. 8, pp. 699-707.

- Hadamard, J.: [1902] Sur les Problèmes aux Dérivées Partielles et leur Signification Physique. *Bull. Univ.*, Princeton, vol. 13.
- Herglotz, G.: [1970] Über das Benndorfsche Problem der Fortpflanzungs-Geschwindigkeit der Erdbebenstrahlen. *Z. Phys.*, vol. 8, pp. 145-147.
- Herlitz, S. I.: [1963] A method for Computing the Emission Distribution in Cylindrical Light Sources. *Ark. for Fys.*, vol. 23, pp. 571-574.
- Hodgson, P. E.: [1963] The Optical Model of Elastic Scattering. Oxford at the Clarendon Press.
- Holmberg, B.: [1952] A remark on the Uniqueness of the Potential Determined from the Asymptotic Phase. *Nuovo Cimento*, Ser. 9, vol. 9, pp. 597-604.
- Hooshyvar, M. A.: [1970] On the Inverse Scattering Problem at Fixed Energy for Tensor and Spin Orbit Potentials. Ph. D. Thesis, Indiana University, October, 1970. *J. Math. Phys.*
- Hunter, J. D.; and Bates, R. H. T.: [1970] Computation of Scattering from a Class of Bodies of Unrestricted Side. *Jour. of Eng. Math.*, vol. 4, pp. 119-128.
- Hylleraas, E. A.: [1948] Calculation of a Perturbing Central Field of Force from the Elastic Scattering Phase-Shifts. *Phys. Rev.*, vol. 74, pp. 48-51.
- Jobert, N.: [1960] Recherche d'une Méthode Permettant de Remonter de la Dispersion des Ondes de Love à la Structure du Manteau Terrestre. Colloque CNRS "Propagation des Ebranlements dans les Milieux Hétérogènes". [1970] Perturbations de la Vitesse de Groupe des Ondes de Love. *Ann. Geophys.*, vol. 26, pp. 273-278.
- Jost, R.; and Kohn, W.: [1952] Construction of a Potential from a Phase-Shift. *Phys. Rev.*, vol. 87, pp. 977-992.
- Kac, I. S.: [1965] *Dokl. Akad. Nauk SSSR Tom 164 n° 5*.
- Kagiwada, H. H.; and Kalaba, R. E.: [1968] Direct and Inverse Problems for Integral Equations Via Initial Value Methods. *Slam, AMS Proceedings*, vol. 1.
- Kay, I.: [1961] Some Remarks Concerning the Bremmer Series. *Jour. of Math. Anal. and Appl.*, vol. 3, pp. 40-49.
- Kay, I.: [1971] The Inverse Scattering Problem for Transmission Line. Workshop on the Mathematics of Profile Inversion. See also Kay, I. and Moses, H. E. The determination of the Scattering Potential from the Spectral Measure Function. *Nuovo Cimento*, vol. 2, p. 917, vol. 3, p. 66, vol. 3, p. 276. Suppl. vol. 5, p. 230.
- Keilis-Borok, V. I.; and Yanovskaja, T. B.: [1967] Inverse Problems of Seismology. *Geophys. J. R. Astr. Soc.*, vol. 13, pp. 223-234.
- Keller, J. B.; Kay, I.; and Shmoyes, J.: [1956] Determination of the Potential from the Scattering Data. *Phys. Rev.*, vol. 102, pp. 557-559.
- Kim, Y. S.; and Vasavada, D. V.: [1966] Perturbed Bound-State Poles in Potential Scattering. *Phys. Rev.*, vol. 150, p. 1236.

- Knopoff, L.: [1960] The Inversion of Surface Wave Dispersion. *Colloque CNRS*, 1960. [1961] Green's Function for Eigenvalue Problems and the Inversion of Love Wave Dispersion Data. *Geophys. J. R. Astr. Soc.*, vol. 4, p. 161. [1962] Higher Order Born Approximations for the Inversion of Love Wave Dispersion Data. *Geophys. J. R. Astr. Soc.*, vol. 7, p. 149.
- Krein, M. G.: [1953] On Some Cases of Effective Determination of the Density of an Inhomogeneous Cord from its Spectral Function. *Dokl. Akad. Nauk SSSR*, vol. 93, pp. 617-621. [1956] *Dokl. Akad. Nauk SSSR*, vol. 111, p. 1167.
- Kunetz, G.; and Rocroi, J. P.: Automatic Processing of Electrical Soundings. *Geophys. Prospecting XVIII*, vol. 2, pp. 157-196.
- Lavrentiev, M. M.: [1967] Some Improperly Posed Problems of Mathematical Physics. Springer-Verlag, Berlin.
- Levinson, N.: [1949a] Determination of the Potential from the Asymptotic Phase. *Phys. Rev.*, vol. 75, p. 1445. [1949b] On the uniqueness of the Potential in a Schrödinger Equation for a Given Asymptotic Phase. *Kgl Danske Vidensk. Selskab. Mat. Fys. Medd.*, vol. 25, 9, pp. 1, 29.
- Levitan, B. M.: [1964] Generalized Translation Operators Israel Program for Scientific Translations. Jerusalem, 1964.
- Levshin, A. L.; Sabitova, T. M.; Valus, V. P.: [1966] Joint Interpretation of Body and Surface Waves Data for a District in Middle Asia. *Geophys. J. R. Astr. Soc.*, vol. 11, pp. 57-66.
- Lions, J. L.: [1968] Contrôle Optimal de Systèmes Gouvernés par des Équations aux Dérivées Partielles. Dunod, Paris.
- Ljance, V. E.: [1964] A Differential Operator with Spectral Singularities, I, II, *Mat. Sb.*, vol. 64 (106), pp. 521-561; vol. 65 (107), pp. 47-103. [*Amer. Math. Soc. Transl.*, Ser. 2, vol. 60, pp. 185-283].
- Ljance, V. E.: [1966] The Inverse Problem for a Nonselfadjoint Operator. *Doklady Akad. Nauk SSSR*, vol. 166, pp. 30-33. [*Soviet Math.*, vol. 7, pp. 27-30].
- Loeffel, J. J.: [1968] On an Inverse Problem in Potential Scattering Theory. *Ann. Inst. H. Poincaré*, vol. 8, pp. 339-347.
- Mal; and Knopoff, L.: [1968] *Journ. of Math. Anal. Appl.*, vol. 21, pp. 431-444.
- Maldonado, C. D.; and Olsen, H. N.: [1965] New Method for Obtaining Emission Coefficients from Emitted Spectral Intensities. Part I, *Journ. Opt. Soc. Amer.*, vol. 55, pp. 17-1953.
- Marcenko, V. A.: [1955] *Dokl. Akad. Nauk SSSR*, vol. 104, p. 695.
- Martin, A.: [1961] *Nuovo Cim.*, vol. 19, p. 1257.
- Martin, A.: [1969] Construction of the Scattering Amplitude from the Differential Cross Sections. *Nuovo Cimento*, vol. 59 A, pp. 131-152.
- Martin A.; and Targonski, Gy.: [1961] On the Uniqueness of a Potential Fitting a Scattering Amplitude at a Given Energy. *Nuovo Cimento*, vol. 20, pp. 1182-1190.

- Mazuel, G. P.: [1968] Review and Extension of Inversion Techniques in Radiation Transfer. Report AA & ES 67-13, Purdue University.
- Melnikov, V. K.: [1959] On Approximate Methods in the Inverse Problem of the Quantum Theory of Scattering. *Usp. Mat. Nauk (N. S.)*, vol. 14, No. 4 (88), pp. 121-131. [*Amer. Math. Soc. Transl.*, Ser. 2, vol. 25, pp. 271-282].
- Miller, W. H.: [1969] WKB Solution of Inversion Problems for Potential Scattering. *J. Chem. Phys.*, vol. 51, pp. 3631-3638.
- Mills, R. L.; Reading, J. F.: [1969] Inversion Problem with Separable Potentials. *Journ. Math. Phys.*, vol. 10, p. 321.
- Minerbo, G. N., and Levy, M. E.: Inversion of Abel's Integral Equation by Means of Orthogonal Polynomials. *Siam J. Numer. Anal.*, vol. 6, pp. 598-616 (1969).
- Mireles, R.: [1966] The Inverse Problem of Electromagnetic Scattering Theory, A Uniqueness Theorem for Cylinders. *J. Math. and Phys.*, vol. 45, pp. 179-187 (1966).
- Naimark, M. A.: [1952] On the Spectrum of a Nonselfadjoint Differential Operator of the Second Order. *Dokl. Akad. Nauk SSSR*, vol. 85, pp. 41-44.
- Nestor, O. H.; and Olsen, H. N.: [1960] Numerical Methods for Reducing Line and Surface Probe Data. *Siam Review*, vol. 2, pp. 200-207.
- Newton, R. G.: [1955] *Phys. Rev.*, vol. 100, p. 412. [1956] Remarks on Scattering Theory. *Phys. Rev.*, vol. 101, pp. 1588-1596. [1962] Construction of Potentials from the Phase Shifts at Fixed Energy. *Journ. of Math. Phys.*, vol. 3, pp. 75-82. [1968] Determination of the Amplitude from the Differential Cross Section by Unitarity. *J. Math. Phys.*, vol. 9, pp. 2050-2055.
- Percival, I. C.: *Proceeding of Physical Society of London*, vol. 80, p. 1290.
- Portinari, J. C.: [1966] Ph. D. Thesis. Massachusetts Institute of Technology.
- Prats, F.; and Toll, J. S.: [1959] Construction of the Dirac Equation Central Potential from Phase Shifts and Bound States. *Phys. Rev.*, vol. 113, pp. 363-370.
- Press, F.: [1968] *Science N. Y.*, vol. 160, pp. 1218-1221.
- Prosser, R. T.: Formal Solutions of Inverse Problems. *Journ. of Math. Phys.*, vol. 10, p. 1819.
- Regge, T.: [1959] Introduction to Complex Orbital Momenta. *Nuovo Cimento*, vol. 14, pp. 951-976.
- Roberts, M. J.: [1964] *Proc. Phys. Soc. (London)*, vol. 84, p. 825.
- Rydbeck, O. E.: [1942] A Theoretical Survey of the Possibilities of Determining the Distribution of the Free Electrons in the Upper Atmosphere. *Trans. Chalmers Univ. Tech.*, Gothebur, Sweden, vol. 1.
- Sabatier, P. C.: [1966a] Le Problème Inverse à énergie fixée en Mécanique Quantique. 222p. Thèse de Doctorat d'Etat: Université Paris-Orsay, Mars, 1966. [1966b] Asymptotic Properties of the Potentials in the Inverse-Scattering Problem at Fixed Energy. *J. Math. Phys.*, vol. 7, pp. 1515-1531. [1966c] Analytic Properties of a Class of Potentials and the Corresponding Jost Functions. *J. Math. Phys.*, vol. 7, pp. 2079-2091. [1967a] General Method for the Inverse Scattering Problem at Fixed Energy. *J. Math.*

- Phys.*, vol. 8, pp. 905-918. [1967b] Interpolation Formulas in the Angular Momentum Plane. *J. Math. Phys.*, vol. 8, pp. 1957-1971. [1968] Approach to Scattering Problems Through Interpolation Formulas and Application to Spin-Orbit Potentials. *J. Math. Phys.*, vol. 9, pp. 1241-1258. [1970] A New Tool for Scattering Studies, vol. 12, pp. 1303-1326. *Journal of Mathematical Physics*. [1972] Complete Solution of the Inverse Problem at Fixed Energy. *Journ. of Math. Phys.*, vol. 13, pp. 675-700.
- Sabatier, P. C.; and Quyen Van Phu, F.: [1971] Numerical Computations in the Inverse Scattering Problem at Fixed Energy. *Phys. Rev.*, D vol. 4, pp 127-132.
- Shilov, G. Ye.: [1965] *Mathematical Analysis. A special course.* Pergamon.
- Shmoys, J.; and Pirraglia, J.: [1971] Diagnostics of Non-Radially Stratified Media. Workshop on Inverse Problems.
- Slichter, L. B.: [1932] The Theory of the Interpretation of Seismic Travel-Time Curves in Horizontal Structures. *Physics*, vol. 3, pp. 273-295.
- Strand, O. N.; and Westwater, E. R.: [1968a] Minimum RMS Estimation of the Numerical Solution of a Fredholm Integral Equation of the 1st kind. *Siam J. Num. An.*, vol. 5, p. 287-295. [1968b] Statistical Estimation of the Numerical Solution of a Fredholm Integral Equation of the 1st Kind. *Journ. Assoc. Comp. Mach.*, vol. 15, 100-114. [1968c] Statistical Information Content of Radiation Measurements Used in Indirect Sensing. *Journ. Atm. Sci.*, vol. 25, pp. 750-758.
- Tabekin, F.: Inverse Scattering Problem for Separable Potentials. *Phys. Rev.*, vol. 177, pp. 1443-1451 (1969).
- Titchmarsh: [1940] *Eigenfunction Expansions Associated with Second Order Differential Equations (Part I).* 1958 Part II. Oxford Clarendon Press.
- Verde, M.: [1958-1959] *Nuclear Physics*, vol. 9, p. 255.
- Verreault, F.: [1965] Théorie des Perturbations d'un Modèle de Terre. *Ann. de Geophys.*, vol. 21, pp. 252-264.
- Vollmer, G.: [1969] Inverse Problem in Atom Scattering in WKB Approach. *Z. Physik*, vol. 226, pp. 423-434.
- Weston, V. H.; and Boerner, N. M.: [1967] Inverse Scattering Investigation. ESD-TR 68-215, Michigan University.
- Westwater, E. R.: [1970] Ground-Based Determination of Temperature Profile by Microwaves. Ph. D. Thesis. Department of Physics and Astrophysics, University of Colorado, 1970.
- Wheeler, J. A.: [1955] *Phys. Rev.*, vol. 99, p. 630.
- Wiechert, E.: [1907] Über Erdbebenwellen I. *Nachr. Ges. Wiss. Göttingen Math. Phys.*, K1, pp. 415-429.
- Yanovskaja, T. B.; et Asbel, I. Y.: [1963] The Determination of Velocities in the Upper Mantle from the Observation on P-Waves. *Geophys. J. R. Astr. Soc.*, vol. 8, p. 3.

ADDITIONAL COMMENTS CONCERNING INVERSION TECHNIQUES

J. Shmoys

Polytechnic Institute of Brooklyn, Long Island Graduate Center,
Farmingdale, New York 11735

ABSTRACT

N73-11630

The chronological development of multidimensional inversion techniques is reviewed briefly. Some current work in this area is described.

As we have seen, the one-dimensional inverse problem has received a great deal of attention over the past 65 years. The literature is tremendous and the effort has extended over a great many areas of application as well as the basic mathematic procedures involved. By comparison the extension to problems in more than one dimension has received relatively little attention. The history of the multidimensional inverse problem differs considerably from that of the one-dimensional one. In the one-dimensional problem, "classical" methods came first, various numerical and modeling procedures were being developed simultaneously with perturbation methods and other approximate procedures for the inverse "wave" problem; finally, exact methods for the inversion of the wave problem were devised.

The multidimensional inverse problem developed in reverse order. Exact inversion methods have been known for almost 20 years [Moses, 1956]. Classical methods using variation of parameters or perturbations have been used in recent years, but relatively sparsely [Cain *et al.*, 1966; Gross and Pirraglia, 1971]. There is as yet no systematic procedure for classical inversion of a multidimensional scattering problem. One might ask, why bother with classical methods if wave methods are available? First, the available exact wave methods appear to depend on Born-type expansions and may therefore converge slowly when the phase shifts are large and strongly dependent on frequency and direction of incidence due to large extent of the unknown medium. Second, classical formulation generally results in a simpler problem; this, at least is the case in the one-dimensional inverse problem. The latter consideration may be rephrased to say that the classical problem requires a different type of data, a type that is sometimes easier to obtain experimentally. For example, the time delay of a wave packet may be easier to measure than the phase shift of the carrier.

In their important paper, Iwata and Nagata [1970] present a perturbation treatment of the eikonal equation. The resulting system of equations, each equation containing terms involving the solutions of all the preceding equations, can be solved. Unfortunately, rather than expand each perturbation in a complete set of functions, they assume a finite series, and the coefficients are obtained from a finite set of observations. The method can probably be carried further, and conditions on the existence of solutions and convergence of the series should be established.

Lavrentiev *et al.* [1970] have provided a fairly systematic presentation of multidimensional inverse problems. The stepping stone for their monograph is the determination of a function of n variables from the values of its mean over a set surfaces. The surfaces constitute an n -parameter family, and the mean must be known as a function of all the parameters. This basic inversion problem in integral geometry is treated carefully, with conditions established for the unknown function as well as for the set of surfaces. Various inverse problems of partial differential equations are reduced to this basic problem of integral geometry. In all the problems treated, the spatial domain of the partial differential equation is the half space, and the data available are assumed to be the response to an impulse source located at a point on the boundary as observed at another point on the boundary.

Lavrentiev et al. deal with the Klein-Gordon type equation, with variable coefficients, giving results that would be applicable directly to plasma diagnostics. The unknown coefficient function is taken to be a perturbation of a constant. This perturbation can be calculated from the response as a function of the coordinates of the source and observation points.

In dealing with the non-dispersive scalar wave equation with variable refractive index is dealt with, *Lavrentiev et al.* immediately restrict the information required to the classical variable, time of arrival of the wave front. Again the problem is considered as a perturbation problem, this time of a one-dimensional index variation which is assumed known. The time delay as a function of source and observation point locations can be inverted to yield the perturbation in refractive index. Other applications of this approach are given.

Clearly, while the multidimensional inverse problem is finally receiving some attention, we are not yet at the end of the road. The utilization of temporal response to an impulse both in the multidimensional as well as in the one-dimensional case has interesting possibilities. Perturbation procedures might be useful if they can be shown to be sufficiently rapidly convergent for reasonable amplitudes of the perturbation.

REFERENCES

- Cain, D. L.; Kliore, A. J.; and Levy, G. S.: The Mariner 4 Occultation Experiment. *AIAA*, Paper No. 66-148, 1966.
- Gross, S. H.; and Pirraglia, J. A.: Errors Incurred in Profile Reconstruction and Methods for Increasing Inversion Accuracies for Occultation Type Measurements, these Proceedings.
- Iwata, K.; Nagata, R.: Calculation of Three-Dimensional Refractive Index Distribution from Interferograms. *J. Opt. Soc. Am.*, vol. 60, p. 133, 1970.
- Lavrentiev; Romanov; and Vasiliev: Multidimensional Inverse Problems. *Lecture Notes in Mathematics*, No. 167, Springer Verlag, Berlin, 1970.
- Moses, H. E.: Calculation of Scattering Potentials from Scattering Data. *Phys. Rev.*, vol. 102, p. 559, 1956.

APPENDIX
CONFERENCE ATTENDEES

Bojarski, F.
16 Circle Drive
Moorestown, New Jersey 08057

Bojarski, N.
16 Circle Drive
Moorestown, New Jersey 08057

Calogero, F.
Istituto di Fisica dell' Universita
Piazza delle Scienze
Rome, Italy

Chadan, K.
Physics Theorique et Hautes Energies
Bat. 211, Faculte Sci
91 Orsay, France

Chahine, M.
Jet Propulsion Laboratory
4800 Oak Grove Drive
Pasadena, California 91103

Chan, K. L.
NASA-Ames Research Center
M.S. 245-4
Moffett Field, California 94035

Chang, H. T.
CIRES
University of Colorado
Boulder, Colorado 80302

Chen, K. H.
Mathematics Department
Stanford University
Stanford, California 94305

Ciftan, M.
Department of the Army
U.S. Army Research Office
Box CM, Duke Street, Durham. No. Car. 27706

Claerbout, J.
Stanford University
Stanford, California 94305

Colin, L.
NASA-AMES Research Center
M.S. 245-1
Moffett Field, California 94035

Collis, R.
Stanford Research Institute
Ravenwood Avenue
Menlo Park, California 94025

Conrath, B. J.
Code 622
NASA-Goddard Space Flight Center
Greenbelt, Maryland 20771

Corbella, G. D.
IMAF, Universidad Nacional de Cordoba
Laprida 854
Cordoba, Argentina

Cox, J.
Department of Physics
Florida Atlantic University
Boca Raton, Florida 33432

Crane, R. K.
MIT Lincoln Labs
Lexington, Massachusetts 02173

Croft, T. A.
Stanford Electronics Labs
Stanford University
Stanford, California 94305

Dahlen, F. A.
Princeton University
Princeton, New Jersey 08540

Degasperis, A.
Istituto di Fisica "G. Marconi"
Universita di Roma
Piazzale Delle Scienze, 5
Roma, Italy

De Santo, J. A.
Naval Research Lab, Code 8172
Washington, D. C. 20390

Dziewonski, A.
University of Texas at Dallas
Dallas, Texas 75230

Eshleman, V.
Stanford Electronic Labs
Stanford University
Stanford, California 94305

Falcone, Jr., V. J.
Air Force Cambridge Research Labs
Laurence G. Hanscom Field
Bedford, Massachusetts 01730

Fleming, H.
National Oceanic and Atmospheric Administration
Washington, D.C. 20023

Forance, E. T.
Office of Naval Research
Department of the Navy
Pasadena, California 91101

Gilbert, F.
University of California at San Diego
La Jolla, California 92037

Gille, J.
Department of Meteorology
Florida State University
Tallahassee, Florida 32306

Gopinath, B.
Bell Telephone Labs
Murray Hill, New Jersey 07974

Gross, S. H.
Polytechnic Institute of Brooklyn
Graduate Center, Route 110
Farmingdale, New York 11735

Grossi, M. D.
Raytheon Company
Sudbury Engineering Facility
528 Boston Post Road
Sudbury, Massachusetts 01776

Howland, J. S.
Department of Mathematics
Stanford University
Stanford, California 94305

Hunt, G.
Space Science Division
4800 Oak Grove Drive
Pasadena, California 91103

Jackson, J. E.
Code 625
NASA-Goddard Space Flight Center
Greenbelt, Maryland 20771

Johnson, L.
University of California at San Diego
La Jolla, California 92037

Jordan, T.
California Institute of Technology
Pasadena, California 91103

Kay, I. W.
Department of Electrical Engineering
Wayne State University
Detroit, Michigan 48202

King, J. I. F.
Air Force Cambridge Research Labs
Laurence G. Hanscom Field
Bedford, Massachusetts 01730

Kliore, A.
Jet Propulsion Laboratory
4800 Oak Grove Drive
Pasadena, California 91103

Lax, P. D.
Courant Institute
New York University
251 Mercer Street
New York, New York 10012

Lockwood, G. E. K.
Communications Research Center
Shirley Bay
Ottawa, Canada

Lusignan, B.
Department of Electrical Engineering
Stanford University
Stanford, California 94305

Marshall, J.
ESL, Incorporated
495 Java Drive
Sunnyvale, California 94086

Mateer, C.
Canadian Meteorological Service
315 Bloor Street West
Toronto, Canada

McAllister, L. G.
Australian Defense Scientific Service
Department of Supply
Weapons Research Establishment
Adelaide, So. Australia

McClatchey, R.
Air Force Cambridge Research Lab
Laurence G. Hanscom Field
Bedford, Massachusetts 01730

McCormick, P. T.
University of Santa Clara
Santa Clara, California 95053

McCulley, L. G.
Informatics Inc.
NASA-Ames Research Center 245-1
Moffett Field, California 94035

Minster, B.
Seismological Lab
Bin 2 - Arroyo Annex
Pasadena, California 91103

Mitra, R.
University of Illinois
Urbana, Illinois 61801

Mochizuki, K.
Mathematics Department
Stanford University
Stanford, California 94305

Moses, H. E.
Air Force Cambridge Research Labs
Laurence G. Hanscom Field
Bedford, Massachusetts 01730

Mulcahy, M.
Department of Geophysical and Geological
Sciences
Guyot Hall
Princeton University
Princeton, New Jersey 08540

Newton, R. G.
Department of Physics
Indiana University
Bloomington, Indiana 47401

Paris, J.
Lockheed Electronics Co.
Houston, Texas 77001

Parker, R.
University of California at San Diego
La Jolla, California 92037

Parry, H. D.
NOAA-National Weather Service W442
8060 - 13th Street
Silver Springs, Maryland 20910

Paul, A. K.
(NOAA)
78 Frieburg
Kronenstr 13
West Germany

Peterson, A. M.
Electronics Research Labs
Stanford University
Stanford, California 94305

Phillips, R. S.
Department of Mathematics
Stanford University
Stanford, California 94305

Phinney, R. A.
Department of Geological and Geophysical
Services
Guyot Hall
Princeton University
Princeton, New Jersey 08540

Portinari, J. C.
Institute de Matematica Pontifica
Universidade Catolica
Rua Marquis de Sao Vicente 233zc-20
Rio de Janeiro, Brazil

Prosser, R. T.
Department of Mathematics
Dartmouth University
Hanover, New Hampshire 03755

Reinisch, B. W.
Lowell Technological Institute Research
Foundation
450 Aiken Street
Lowell, Massachusetts 01854

Revah, I.
Department R-S-R
Centre National d' Etudes des
Telecommunications
3 Av. de la Republique, 92 Issy
Les Moulineaux, France

Reynolds, R.
NASA-Ames Research Center
M.S. 245-4
Moffett Field, California 94035

Richards, P.
IGPP Laboratory
University of California at San Diego
San Diego, California 92037

Richter, J. H.
Naval Electronics Labs
San Diego, California 92152

Rodgers, C. D.
Clarendon Laboratory
Oxford University
Oxford, England OX1-3PU

Tabatier, P.
Institute of Mathematical Sciences
University of Montpellier
Montpellier, France

Sarason, L.
Mathematics Department
Stanford University
Stanford, California 94305

Schubert, G.
Planetary and Space Sciences Department
University of California at Los Angeles
Los Angeles, California 90024

Schwartz, K.
American Nucleonics Corp.
6036 Varie Avenue
Woodland Hills, California 91364

Sharpe, C. S.
Department of Electrical Engineering
University of Michigan
Ann Arbor, Michigan 48104

Shellman, C. H.
Naval Electronics Laboratory
271 Catalina Blvd.
San Diego, California 92152

Shmoys, J.
Polytechnic Institute of Brooklyn
Graduate Center, Route 110
Farmingdale, New York 11735

Smith, G.
Stanford Research Institute
Ravenswood Avenue
Menlo Park, California 94025

Smith, W.
Suite 300
National Oceanic and Atmospheric
Administration
3737 Branch Avenue
Washington, D. C. 20031

Soicher, H.
Institute for Exploratory Research
U.S. Army Electronics Command
Fort Monmouth, New Jersey 07703

Sondhi, M. M.
Bell Telephone Labs
Murray Hill, New Jersey 07974

Stewart, R. W.
Goddard Institute for Space Studies
2880 Broadway
New York, New York 10025

St. Germain, R.
Raytheon Co. Equip. Div.
528 Boston Post Road
Sudbury, Massachusetts 01776

Strand, O.
National Oceanic and Atmospheric Administration
Boulder, Colorado 80302

Taylor, F.
Bldg. 183-301
Jet Propulsion Laboratory
4800 Oak Grove Drive
Pasadena, California 91103

Thompson, K. W.
Department of Physics
Florida Atlantic University
Boca Raton, Florida 33432

Twersky, V.
Mathematics Department
University of Illinois
Chicago, Illinois 60680

Ungar, S.
Stanford Electronics Labs
Stanford University
Stanford, California 94305

Uthe, E. E.
Stanford Research Institute
Ravenswood Avenue
Menlo Park, California 94025

Vaillancourt, R.
Department of Mathematics
Stanford University
Stanford, California 94305

Wallio, A.
152-A
NASA-Langley Research Center
Hampton, Virginia 23369

Waters, J.
Room 26-650
Research Lab of Electronics
Massachusetts Institute of Technology
Cambridge, Massachusetts 02139

Watt, T.
Stanford Research Institute
Ravenswood Avenue
Menlo Park, California 94025

Weston, V. H.
Department of Mathematical Sciences
Purdue University
Lafayette, Indiana 47407

Westwater, E. R.
National Oceanic and Atmospheric
Administration
Boulder, Colorado 80302

Whitten, R. C.
NASA-Ames Research Center
M.S. 245-4
Moffett Field, California 94035

Wiggins, R.
Department of Physics
University of Toronto
Toronto 5, Ontario, Canada

Wright, J. W.
National Oceanic and Atmospheric
Administration
Boulder, Colorado 80302